**PCT** WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT C

(51) International Patent Classification 6 : C12Q 1/68 — A1

(11) International Publication Number: **WO 95/21944**

(43) International Publication Date: 17 August 1995 (17.08.95)

(21) International Application Number: PCT/US95/01863

(22) International Filing Date: 14 February 1995 (14.02.95)

(30) Priority Data:
08/195,485 14 February 1994 (14.02.94) US

(60) Parent Application or Grant
(63) Related by Continuation
US 08/195,485 (CIP)
Filed on 14 February 1994 (14.02.94)

(71) Applicant *(for all designated States except US)*: SMITHKLINE BEECHAM CORPORATION [US/US]; Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: ROSENBERG, Martin [US/US]; 241 Mingo Road, Royersford, PA 19468 (US). DEBOUCK, Christine [BE/US]; 667 Pugh Road, Wayne, PA 19087 (US). BERGSMA, Derk [US/US]; 271 Irish Road, Berwyn, PA 19312 (US).

(74) Agents: JERVIS, Herbert, H. et al.; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US).

(81) Designated States: JP, US, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

**Published**
*With international search report.*

(54) Title: DIFFERENTIALLY EXPRESSED GENES IN HEALTHY AND DISEASED SUBJECTS

(57) Abstract

The present invention involves methods and compositions for identifying genes which are differentially expressed in a normal healthy animal and an animal having a selected disease or infection, and methods for diagnosing diseases or infections characterized by the presence of those genes, despite the absence of knowledge about the gene or its function. The methods involve the use of a composition suitable for use in hybridization which consists of a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences for hybridization. Each sequence comprises a fragment of an EST isolated from an identified DNA library prepared from tissue or cell samples of a healthy animal, an animal with a selected disease or infection, and any combination thereof. Differences in hybridization patterns produced through use of this composition and the specified methods enable diagnosis of disease based on differential expression of genes of unknown function, and enable the identification of those genes and the proteins encoded thereby.

*FOR THE PURPOSES OF INFORMATION ONLY*

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

| | | | | | |
|---|---|---|---|---|---|
| **AT** | Austria | **GB** | United Kingdom | **MR** | Mauritania |
| **AU** | Australia | **GE** | Georgia | **MW** | Malawi |
| **BB** | Barbados | **GN** | Guinea | **NE** | Niger |
| **BE** | Belgium | **GR** | Greece | **NL** | Netherlands |
| **BF** | Burkina Faso | **HU** | Hungary | **NO** | Norway |
| **BG** | Bulgaria | **IE** | Ireland | **NZ** | New Zealand |
| **BJ** | Benin | **IT** | Italy | **PL** | Poland |
| **BR** | Brazil | **JP** | Japan | **PT** | Portugal |
| **BY** | Belarus | **KE** | Kenya | **RO** | Romania |
| **CA** | Canada | **KG** | Kyrgystan | **RU** | Russian Federation |
| **CF** | Central African Republic | **KP** | Democratic People's Republic of Korea | **SD** | Sudan |
| **CG** | Congo | | | **SE** | Sweden |
| **CH** | Switzerland | **KR** | Republic of Korea | **SI** | Slovenia |
| **CI** | Côte d'Ivoire | **KZ** | Kazakhstan | **SK** | Slovakia |
| **CM** | Cameroon | **LI** | Liechtenstein | **SN** | Senegal |
| **CN** | China | **LK** | Sri Lanka | **TD** | Chad |
| **CS** | Czechoslovakia | **LU** | Luxembourg | **TG** | Togo |
| **CZ** | Czech Republic | **LV** | Latvia | **TJ** | Tajikistan |
| **DE** | Germany | **MC** | Monaco | **TT** | Trinidad and Tobago |
| **DK** | Denmark | **MD** | Republic of Moldova | **UA** | Ukraine |
| **ES** | Spain | **MG** | Madagascar | **US** | United States of America |
| **FI** | Finland | **ML** | Mali | **UZ** | Uzbekistan |
| **FR** | France | **MN** | Mongolia | **VN** | Viet Nam |
| **GA** | Gabon | | | | |

differentially expressed genes in healthy and diseased subjects

Cross Reference to Related Applications:

This application is a continuation-in-part application of U.S. Serial No. 08/195,485 filed February 14, 1994, the contents of which are incorporated herein by reference.

Field of the Invention

The present invention relates to the use of immobilized oligonucleotide/polynucleotide or polynucleotide sequences for the identification, sequencing and characterization of genes which are implicated in disease, infection, or development and the use of such identified genes and the proteins encoded thereby in diagnosis, prognosis, therapy and drug discovery.

Background of the Invention

Identification, sequencing and characterization of genes, especially human genes, is a major goal of modern scientific research. By identifying genes, determining their sequences and characterizing their biological function, it is possible to employ recobinant DNA technology to produce large quantities of valuable "gene products", e.g., proteins and peptides. Additionally, knowledge of gene sequences can provide a key to diagnosis, prognosis and treatment of a variety of disease states in plants and animals which are characterized by inappropriate expression and/or repression of selected gene(s) or by the influence of external factors, e.g., carcinogens or teratogens, on gene function. The term disease-associated genes(s) is used herein in its broadest sence to mean not only genes associated with classical inherited diseases, but also those associated with genetic predisposition to disease as well as infectious or pathogenic states resulting from gene expression by infectious agents or the effect on host cell gene expression by the presence of such a pathogen or its products Locating disease-associated genes will permit the development of diagnostic and prognostic reagents and methods, as well as possible therapeutic regimens, and the discovery of new drugs for treating or preventing the occurrence of such diseases.

Methods have been described for the identification of certain novel gene sequences, referred to as Expressed Sequence Tags (EST) [see, e.g., Adams et al, Science, 252:1651-1656 (1991); and International Patent Application No. WO93/00353, published January 7, 1993]. Conventially, an EST is a specific cDNA polynucleotide sequence, or tag, about 150 to 400 nucleotides in length, derived from

a messenger RNA molecule by reverse transcription, which is a marker for, and component of, a human gene actually transcribed *in vivo*. However, as used herein an EST also refers to a genomic DNA fragment derived from an organism, such as a microorganism,the DNA of which lacks intron regions.

A variety of techniques have been described for identifying particular gene sequences on the basis of their gene products. For example, several techniques are described in the art [see, e.g., International Patent Application No. WO91/07087, published May 30, 1991]. Additionally, known methods exist for the amplification of desired sequences [see, e.g., International Patent Application No. WO91/17271, published November 14, 1991, among others].

However, at present, there exist no established methods for filling the need in the art for methods and reagents which employ fragments of differentially expressed genes of known, unknown (or previously unrecognized ) function or consequence to provide diagnostic and therapeutic methods and reagents for diagnosis and treatment of disease or infection, which conditions are characterized by such genes and gene products. It should be appreciated that it is the expression differences that are diagnostic of the altered state (e.g., predisease, disease, pathogenic, progression or infectious). Such genes associated with the altered state are likely to be the targets of drug discovery, whether the genes are the cause or the effect of the condition, identification of such genes provides insight into which gene expression needs to be re-altered in order to reestablished the healthy state.

## Summary of the Invention

In one aspect, the invention provides methods for identifying gene(s) which are differentially expressed, for example, in a normal healthy organism and an organism having a disease. The method involves producing and comparing hybridization patterns formed between samples of expressed mRNA or cDNA polynucleotide sequences obtained from either analogous cells, tissues or organs of a healthy organism and a diseased organism and a defined set of oligonucleotide/polynucleotide/polynucleotide sequence probes from either an healthy organism or a diseased organism immobilized on a support. Those defined oligonucleotide/polynucleotide sequences are representative of the total expressed genetic component of the cells, tissues, organs or organism as defined the collection of partial cDNA sequences (ESTs). The differences between the hybridization patterns permit identification of those particular EST or gene-specific oligonucleotide/polynucleotide sequences associated with differential expression, and the identification of the EST permits identification of the clone from which it was

derived and using ordinary skill further cloning and, if desired, sequencing of the full-length cDNA and genomic counterpart, i.e., gene, from which it was obtained.

In another aspect, the invention provides methods substantially similar to those described above, but which permit identification of those gene(s) of a pathogen which are expressed in any biological sample of an infected organism based on comparative hybridization of RNA/cDNA samples derived from a healthy versus infected organism, hybridized to an oligonucleotide/polynucleotide set representative of the gene coding complement of the pathogen of interest.

In another aspect, the invention provides methods substantially similar to those described above, but which permit identification of those ESTs-specific oligonucleotide/polynucleotide sequences of host gene(s) which represent genes being differentially expressed/ altered in expression by the disease state, or infection and are expressed in any biological sample of an infected organism based on comparative hybridization of RNA/cDNA samples derived from a healthy versus infected organism of interest.

In a further aspect, the methods described above and in detail below, also provide methods for diagnosis of diseases or infections characterized by differentially expressed genes, the expression of which has been altered as a result of infection by the pathogen or disease causing agent in question. All identified differences provide the basis for diagnostic testing be it the altered expression of endogenous genes or the patterned expression of the genes of the infecting organism. Such patterns of altered expression are defined by comparing RNA/cDNA from the two states hybridized against a panel of oligonucleotide/polynucleotides representing the expressed gene component of a cell, tissue, organ or organism as defined by its collection of ESTs.

Yet a further aspect of this invention provides a composition suitable for use in hybridization, which comprises a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences for hybridization, each sequence comprising a fragment of an EST isolated from a cDNA or DNA library prepared from at least one selected tissue or cell sample of a healthy (i.e., pre-disease state) animal, at least one analogous sample of an animal having a disease, at least one analogous sample of an animal infected with a pathogen or the pathogen itself, or any combination or multiple combinations thereof.

An additional aspect of the invention provides an isolated gene sequence which is differentially expressed in a normal healthy animal and an animal having a disease, and is identified by the methods above. Similarly, an isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal can be identified by the methods above.

Yet another aspect of the invention is that it provides not only a means for a static diagnostic but also provides a means for a carrying out the procedure over time to measure disease progression as well as monitoring the efficacy of disease treatment regimes including an toxicological effects thereof.

Another aspect of the invention is an isolated protein produced by expression of the gene sequences identified above. Such proteins are useful in therapeutic compositions or diagnostic compositions, or as targets for drug development.

Other aspects and advantages of the present invention are described further in the following detailed description of the preferred embodiments thereof.

Detailed Description of the Invention

The present invention meets the unfulfilled needs in the art by providing methods for the identification and use of gene fragments and genes, even those of unknown full length sequence and unknown function, which are differentially expressed in a healthy animal and in an animal having a specific disease or infection by use of ESTs derived from DNA libraries of healthy and/or diseased/infected animals. Employing the methods of this invention permits the resulting identification and isolation of such genes by using their corresponding ESTs and thereby also permits the production of protein products encoded by such genes. The genes themselves and/or protein products, if desired, may be employed in the diagnosis or therapy of the disease or infection with which the genes are associated and in the development of new drugs therefor.

It has been appreciated that one or more differentially identified EST or gene-specific oligonucleotide/polynucleotides define a pattern of differentially expressed genes diagnostic of a predisease, disease or infective state. A knowledge of the specific biological function of the EST is not required only that the ESTs identifies a gene or genes whose altered expression is associated reproducibly with the predisease, disease or infectious state. The differences permit the identification of gene products altered in their expression by the disease and represent those products most likely to be targets of therapeutic intervention. Similarly, the product may be of the infecting organism itself and also be an effective target of intervention.

*1. Definitions.*

Several words and phrases used throughout this specification are defined as follows:

As used herein, the term "gene" refers to the genomic nucleotide sequence from which a cDNA sequence is derived, which cDNA produces an EST, as

described below. The term gene classically refers to the genomic sequence, which, upon processing, can produce different cDNAs, e.g., by splicing events. However, for ease of reading, any full-length counterpart cDNA sequence which gives rise to an EST will also be referred to by shorthand herein as a 'gene'.

The term "organism" includes without limitation, microbes, plants and animals.

The term "animal" is used in its broadest sense to include all members of the animal kingdom, including humans. It should be understood, however, that according to this invention the same species of animal which provides the biological sample also is the source of the defined immobilized oligonucleotide/polynucleotides as defined below.

The term "pathogen" is defined herein as any molecule or organism which is capable of infecting an animal or plant and replicating its nucleic acid sequences in the cells or tissues of that animal or plant . Such a pathogen is generally associated with a disease condition in the infected animal or plant. Such pathogens may include viruses, which replicate intra- or extra-cellularly, or other organisms, such as bacteria, fungi or parasites, which generally infect tissues or the blood. Certain pathogens or microorganisms are known to exist in sequential and distinguishable stages of development, e.g., latent stages, infective stages, and stages which cause symptomatic diseases. In these different stages, the pathogens are anticipated to express differentially certain genes and/or turn on or off host cell gene expression.

As used herein, the term "disease" or "disease state" refers to any condition which deviates from a normal or standardized healthy state in an organism of the same species in terms of differential expression of the organism's genes. In other words, a disease state can be any illness or disorder be it of genetic or environmental origin , for example, an inherited disorder such as certain breast cancers, or a disorder which is characterized by expression of gene(s) normally in an inactive, 'turned off' state in a healthy animal, or a disorder which is characterized by under-expression or no expression of gene(s) which is normally activated or 'turned on' in a normal healthy animal. Such differential expression of genes may also be detected in a condition caused by infection, inflammation, or allergy, a condition caused by development or aging of the animal, a condition caused by administration of a drug or exposure of the animal to another agent, e.g., nutrition, which affects gene expression. Essentially, the methods described herein can be adapted to detect differential gene expression resulting from any cause, by manipulation of the defined oligonucleotide/polynucleotides and the samples tested as described below. The

concept of disease or disease state also includes its temporal aspects in terms of progression and treatment.

The phrase "differentially expressed" refers to those situations in which a gene transcript is found in differing numbers of copies, or in activated vs inactivated states, in different cell types or tissue types of an organism, having a selected disease as contrasted to the levels of the gene transcript found in the same cells or tissues of a healthy organism. Genes may be differentially expressed in differing states of activation in microorganisms or pathogens in different stages of development. For example, multiple copies of gene transcripts may be found in an organism having a selected disease, while only one, or significantly fewer copies, of the same gene transcript are found in a healthy organism, or vice-versa.

As used herein, the term "solid support" refers to any known substrate which is useful for the immobilization of large numbers of oligonucleotide/polynucleotide sequences by any available method to enable detectable hybridization of the immobilized oligonucleotide/polynucleotide sequences with other polynucleotide sequences in a sample. Among a number of available solid supports, one desirable example is the supports described in International Patent Application No. WO91/07087, published May 30, 1991.Also useful are suports such as but not limited to nitrocellulose, mylein, glass, silica ans Pall Biodyne C® It is also anticipated that improvements yet to be made to conventional solid supports may also be employed in this invention.

The term "surface" means any generally two-dimensional structure on a solid support to which the desired oligonucleotide/polynucleotide sequence is attached or immobilized. A surface may have steps, ridges, kinks, terraces and the like.

As used herein, the term "predefined region" refers to a localized area on a surface of a solid support on which is immobilized one or multiple copies of a particular oligonucleotide/polynucleotide sequence and which enables the identification of the oligonucleotide/polynucleotide at the position, if hybridization of that oligonucleotide/polynucleotide to a sample polynucleotide occurs.

By "immobilized" refers to the attachment of the oligonucleotide/polynucleotide to the solid support. Means of immobilization are known and conventional to those of skill in the art, and may depend on the type of support being used.

By "EST" or "Expressed Sequence Tag" is meant a partial DNA or cDNA sequence of about 150 to 500, more preferably about 300, sequential nucleotides of a longer sequence obtained from a genomic or cDNA library prepared from a selected cell, cell type, tissue or tissue type, organ or organism which longer

sequence corresponds to an mRNA of a gene found in that library. An EST is generally DNA. One or more libraries made from a single tissue type typically provide at least about 3000 different (i.e., unique) ESTs and potentially the full complement of all possible ESTs representing all cDNAs e.g., 50,000-100,000 in an animal such as a human. Further background and information on the construction of ESTs is described in M. D. Adams et al, Science, 252:1651-1656 (1991); and International Application Number PCT/US92/05222 (January 7, 1993).

As used herein, the term "defined oligonucleotide/polynucleotide sequence" refers to a known nucleotide sequence fragment of a selected EST or gene. This term is used interchangeably with the term "fragments of EST". These sequential sequences are generally comprised of between about 15 to about 45 nucleotides and more preferably between about 20 to about 25 nucleotides in length. Thus any single EST of 300 nucleotides in length may provide about 280 different defined oligonucleotide/polynucleotide sequences of 20 nucleotides in length (e.g., 20-mers). The lengths of the defined oligonucleotide/polynucleotides may be readily increased or decreased as desired or needed, depending on the limitations of the solid support on which they may be immobilized or the requirements of the hybridization conditions to be employed.The length is generally guided by the principle that it should be of sufficient length to insure that it is one average only represented once in the population to be examined. Generally, these defined oligonucleotide/polynucleotides are RNA or DNA and are preferably derived from the anti-sense strand of the EST sequence or from a corresponding mRNA sequence to enable their hybridization with samples of RNA or DNA. Modified nucleotides may be incorporated to increase stability and hybridization properties.

By the term "plurality of defined oligonucleotide/polynucleotide sequences" is meant the following. A surface of a solid support may immobilize a large number of "defined oligonucleotide/polynucleotides". For example, depending upon the nature of the surface, it can immobilize from about 300 to upwards of 60,000 defined 20-mer oligonucleotide/polynucleotides. It is anticipated that future improvements to solid surfaces will permit considerably larger such pluralities to be immobilized on a single surface. A "plurality" of sequences refers to the use on any one solid support of multiple different defined oligonucleotide/polynucleotides from a single EST from a selected library, as well as multiple different defined oligonucleotide/polynucleotides from different ESTs from the same library or many libraries from the same or different tissues, and may also include multiple identical copies of defined oligonucleotide/polynucleotides. Ultimately a pluarality has at least one oligonucleotide/polynucleotide per expressed gene in the entire organism For example, from a library producing about 5,000-10,000 ESTs, a single support can

include at least about 1-20 defined oligonucleotide/polynucleotides representing every EST in that library. The composition of defined oligonucleotide/polynucleotides which make up a surface according to this invention may be selected or designed as desired.

The term "sample" is employed in the description of this invention in several important ways. As used herein, the term "sample" encompasses any cell or tissue from an organism. Any desired cell or tissue type in any desired state may be selected to form a sample. For example, the sample cell desired may be a human T cell; the desired cell type for use in this invention may be a quiescent T cell or an activated T cell.

By the phrase "analogous sample" or "analogous cell or tissue" is meant that according to this invention when the ESTs which provide the defined oligonucleotide/polynucleotides are produced from a cDNA library prepared from a single tissue or cell type source sample, e.g., liver tissue of a human, then the samples used to hybridize to those immobilized defined oligonucleotide/polynucleotides are preferably provided by the same type of sample from either a healthy or diseased animal, i.e., liver tissue of a healthy human and liver tissue of a diseased or infected human or from a human suspected of having that disease or infection. Alternatively, if the surface contains defined oligonucleotide/polynucleotides from multiple cells or tissues, then the "samples" which are hybridized thereto can be but are not limited to samples obtained from analogous multiple tissues or cells.

By the term "detectably hybridizing" means that the sample from the healthy organism or diseased or infected organism is contacted with the defined oligonucleotide/polynucleotides on the surface for sufficient time to permit the formation of patterns of hybridization on the surfaces caused by hybridization between certain polynucleotide sequences in the samples with the certain immobilized defined oligonucleotide/polynucleotides. These patterns are made detectable by the use of available conventional techniques, such as fluorescent labelling of the samples. Preferably hybridization takes place under stringent conditions, e.g., revealing homologies of about 95%. However, if desired, other less stringent conditions may be selected. Techniques and conditions for hybridization at selected stringencies are well known in the art [see, e.g., Sambrook et al, Molecular Cloning, A Laboratory Manual., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1989)].

*II. Compositions of The Invention*

The present invention is based upon the use of ESTs from any desired cell or tissue in known technologies for oligonucleotide/polynucleotide hybridization.

*A. ESTs*

An EST, as defined above, is for an animal, a sequence from a cDNA clone that corresponds to an mRNA. The EST sequences useful in the present invention are isolated preferably from cDNA libraries using a rapid screening and sequencing technique. Custom made cDNA libraries are made using known techniques. See, generally, Sambrook et al, cited above. Briefly, mRNA from a selected cell or tissue is reverse transcribed into complementary DNA (cDNA) using the reverse transcriptase enzyme and made double-stranded using RNase H coupled with DNA polymerase or reverse transcriptase. Restriction enzyme sites are added to the cDNA and it is cloned into a vector. The result is a cDNA library. Alternatively, commercially available cDNA libraries may be used. Libraries of cDNA can also be generated from recombinant expression of genomic DNA using known techniques, including polymerase chain reaction-derived techniques.

ESTs (which can range from about 150 to about 500 nucleotides in length, preferably about 300 nucleotides) can be obtained through sequence analysis from either end of the cDNA insert. Desirably, the DNA libraries used to obtain ESTs use directional cloning methods so that either the 5' end of the cDNA (likely to contain coding sequence) or the 3' end (likely to be a non-coding sequence) can be selectively obtained.

In general, the method for obtaining ESTs comprises applying conventional automated DNA sequencing technology to screen clones, advantageously randomly selected clones, from a cDNA library. The cDNA libraries from the desired tissue can be preprocessed, or edited, by conventional techniques to reduce repeated sequencing of high and intermediate abundance clones and to maximize the chances of finding rare messages from specific cell populations. Preferably, preprocessing includes the use of defined composition prescreening probes, e.g., cDNA corresponding to mitochondria, abundant sequences, ribosomes, actins, myelin basic polypeptides, or any other known high abundance peptide. These prescreening probes used for preprocessing are generally derived from known ESTs. Other useful preprocessing techniques include subtraction hybridization, which preferentially reduces the population of highly represented sequences in the library [e.g., see Fargnoli et al, Anal. Biochem., 187:364 (1990)] and normalization, which results in all sequences being represented in approximately equal proportions in the library [Patanjali et al, Proc. Natl. Acad. Sci. USA, 88:1943 (1991)]. Additional prescreening/differential screening approaches are known to those skilled in the art.

ESTs can then be generated from partial DNA sequencing of the selected clones. The ESTs useful in the present invention are preferably generated using low redundancy of sequencing, typically a single sequencing reaction. While

single sequencing reactions may have an accuracy as low as 90%, this nevertheless provides sufficient fidelity for identification of the sequence and design of PCR primers.

If desired, the location of an EST in a full length cDNA is determined by analyzing the EST for the presence of coding sequence. A conventional computer program is used to predict the extent and orientation of the coding region of a sequence (using all six reading frames). Based on this information, it is possible to infer the presence of start or stop codons within a sequence and whether the sequence is completely coding or completely non-coding or a combination of the two. If start or stop codons are present, then the EST can cover both part of the 5'-untranslated or 3'-untranslated part of the mRNA (respectively) as well as part of the coding sequence. If no coding sequence is present, it is likely that the EST is derived from the 3' untranslated sequence due to its longer length and the fact that most cDNA library construction methods are biased toward the 3' end of the mRNA. It should be understood that both coding and non-coding regions may provide ESTs equally useful in the described invention.

A number of specific ESTs suitable for use in the present invention are described above Adams et al (supra), which may be incorporated by reference herein, to describe non-essential examples of desirable ESTs. Other ESTs exist in the art which may also be useful in this invention, as will ESTs yet to be developed by these known techniques.

*B. Preparing the Solid Support of the Invention*

Oligonucleotide sequences which are fragments of defined sequence are derived from each EST by conventional means, e.g., conventional chemical synthesis or recombinant techniques. Each defined oligonucleotide/polynucleotide sequence as described above is a fragment, can be, but is not necessarily an anti-sense fragment, of an EST isolated from a DNA library prepared from a selected cell or tissue type from a selected animal. For use in the present invention, it is presently preferred that the defined oligonucleotide/polynucleotide sequences are 20-25mers. As described above, for each EST a number of such 20-25mers may be generated. The lengths may vary as described above as well as the composition. For example oligonucleotide/polynucleotides can be modified based on the Oligo 4.0 or simiolar programs to predict hybridization potential or to include modifieid nucleotides for the reasons given above. It is alos appreciated that large DNA segments may be employed including entire ESTs or even full length genes particular when inserted into cloning vectors.

A plurality of these defined oligonucleotide/polynucleotide sequences are then attached to a selected solid support conventionally used for the attachment of nucleotide sequences again by known means. In contrast to other technologies available in the art, this support is designed to contain defined, not random, oligonucleotide/polynucleotide sequences. The EST fragments, or defined oligonucleotide/polynucleotide sequences, immobilized on the solid support can include fragments of one or more ESTs from a library of at least one selected tissue or cell sample of a healthy animal, at least one analogous sample of the animal having a disease, at least one analogous sample of the animal infected with a pathogen, and any combination thereof.

Numerous conventional methods are employed for attaching biological molecules such as oligonucleotide/polynucleotide sequences to surfaces of a variety of solid supports. See, e.g., Affinity Techniques. Enzyme Purification: Part B. Methods in Enzymology, Vol. 34, ed. W.B. Jakoby, M. Wilcheck, Acad. Press, NY (1974); Immobilized Biochemicals and Affinity Chromatography, Advances in Experimental Medicine and Biology, vol. 42, ed. R. Dunlap, Plenum Press, NY (1974); U. S. Patent No. 4,762,881; U. S. Patent No. 4,542,102; European Patent Publication No. 391,608 (October 10, 1990); U. S. Patent No. 4,992,127 (Nov. 21, 1989).

One desirable method for attaching oligonucleotide/polynucleotide sequences derived from ESTs to a solid support is described in International Application No. PCT/US90/06607 (published May 30, 1991). Briefly, this method involves forming predefined regions on a surface of a solidsupport, where the predefined regions are capable of immobilizing ESTs. The methods make use of binding substances attached to the surface which enable selective activation of the predefined regions. Upon activation, these binding substances become capable of binding and immobilizing oligonucleotide/polynucleotides based on EST or longer gene sequences.

Any of the known solid substrates suitable for binding oligonucleotide/polynucleotides at pre-defined regions on the surface thereof for hybridization and methods for attaching the oligonucleotide/polynucleotides thereto may be employed by one of skill in the art according to this invention. Similarly, known conventional methods for making hybridization of the immobilized oligonucleotide/polynucleotides detectable, e.g., fluorescence, radioactivity, photoactivation, biotinylation, solid state circuitry, and the like may be used in this invention.

Thus, by resorting to known techniques, the invention provides a composition suitable for use in hybridization which consists of a surface of a solid

support on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences for hybridization. For example, one composition of this invention is a solid support on which are immobilized oligos of EST fragments from a library constructed from a single cell type, e.g., a human stem cell, or a single tissue, e.g., human liver, from a healthy human. Still another composition of this invention is another solid support on which are immobilized oligos of EST fragments from a library constructed from a single cell type or a tissue from a human having a selected disease or predispositon to a selected disease, e.g., liver cancer.

Another embodiment of the compositions of this invention include a single solid support having oligonucleotides of ESTs from both single cell or single tissue libraries from both a healthy and diseased human. Still other embodiments include a single support on which are immobilized oligos of EST fragments from more than one tissue or cell library from a healthy human or a single support on which are immobilized more than one tissue or cell library from both healthy and diseased animals or humans. A preferred composition of this invention is anticipated to be a single support containing oligos of ESTs for all known cells and tissues from a selected organism.

## *III. The Methods of the Invention*

### *A. Identification of Genes*

The present invention employs the compositions described above in methods for identifying genes which are differentially expressed in a normal healthy organism and an organism having a disease or infection. These methods may be employed to detect such genes, regardless of the state of knowledge about the function of the gene. The method of this invention by use of the compositions containing multiple defined EST fragments from a single gene as described above is able to detect levels of expression of genes or in other cases simply the expression or lack thereof, which differ between normal, healthy organisms and organisms having a selected disease, disorder or infection.

One such method employs a first surface of a solid support on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences, described above, of ESTor longer gene fragment isolated from a cDNA library prepared from at least one selected tissue or cell sample of a healthy animal (the "healthy test surface") and a second such surface on which is immobilized at pre-defined regions a plurality of defined oligonucleotide/polynucleotide sequences of ESTor longer gene fragment isolated from at least one analogous tissue of an animal having a selected disease (the "disease

test surface"). These test surfaces may be standardized for the selected animal or selected cell or tissue sample from that animal (i.e., they are prescreened for polymorphisms in the species population).

Polynucleotide sequences are then isolated from mRNA and/or cDNA from a biological sample from a known healthy animal ("healthy control") and a second sample is similarly prepared from a sample from a known diseased animal ("disease sample"). These two samples are desirably selected from the cell or tissue analogous to that which provided the immobilized oligonucleotide/polynucleotides.

According to the method the healthy control sample is contacted with one set of the healthy test surface and the disease test surface described above for a time sufficient to permit detectable hybridization to occur between the sample and the immobilized defined oligonucleotide/polynucleotides on each surface. The results of this hybridization are a first hybridization pattern formed between the nucleotides of healthy control and the healthy test surface and a second hybridization pattern formed between the nucleotides of healthy control sample and the disease test surface.

In a similar manner, the disease sample is detectably hybridized to another set of healthy test and disease test surfaces, forming a third hybridization pattern between the disease sample and healthy test surface and a fourth hybridization pattern between the disease sample and the disease test surface.

Comparing the four hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding ESTor longer gene fragment from which the oligonucleotide/polynucleotides are obtained.

In another embodiment of the method of this invention, the same process is employed, with the exception that plurality of defined oligonucleotide/polynucleotide sequences forming the healthy test sample and the disease test sample surfaces are immobilized on a single solid support. For example, each fragment of an EST or longer gene fragment on the surface is isolated from at least two cDNA libraries prepared from a selected cell or tissue sample of a healthy animal and an analogous selected cell or tissue sample of an animal having a disease.

According to this embodiment, the healthy control sample is detectably hybridized to a copy of this single solid surface, forming one hybridization pattern with oligonucleotide/polynucleotides associated with both the healthy and diseased animal. Similarly, the disease sample is detectably hybridized to a second

copy of this single solid surface, forming one hybridization pattern with oligonucleotide/polynucleotides associated with both the healthy and diseased animal. Comparing the two hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding ESTor longer gene fragment from which the oligonucleotide/polynucleotides are obtained.

The identification of one or more ESTs as the source of the defined oligonucleotide/polynucleotide which produced a "difference" in hybridization patterns according to these methods permits ready identification of the gene from which those ESTs were derived. Because oligonuleotides are of sufficient length that they will hybridize under stringent conditions only with a RNA/cDNA for that gene to which they correspond, the oligo can be used to identify the EST and in turn the clone from which it was derived and by subsequent cloning, obtain the sequence of the full-length cDNA and its genomic counterparts, i.e., the gene, from which it was obtained.

In other words, the ESTs identified by the method of this invention can be employed to determine the complete sequence of the mRNA, in the form of transcribed cDNA, by using the EST as a probe to identify a cDNA clone corresponding to a full-length transcript, followed by sequencing of that clone. The EST or the full length cDNA clone can also be used as a probe to identify a genomic clone or clones that contain the complete gene including regulatory and promoter regions, exons, and introns.

It should be appreciated that one does not have to be restricted in using ESTs from a particular tissue from which probe RNA or cDNA is obtained, rather any or all ESTs (known or unknown) may be placed on the support. Hybridization will be used a form diagnostic patterns or to identifiy which particular EST is detected. For example, all known ESTs from an organism are used to produce a "master" solid support to which control sample and disease samples are alternately hybridized. One then detects a pattern of hybridization associated with the particular disaease state which then forms the basis of a diagnostic test or the isolation of disease specific ESTs from which the intact gene may be cloned and sequenced leading uiltimately to a defined therapuetic target.

Methods for obtaining complete gene sequences from ESTs are well-known to those of skill in the art. See, generally, Sambrook et al, cited above. Briefly, one suitable method involves purifying the DNA from the clone that was

sequenced to give the EST and labeling the isolated insert DNA. Suitable labeling systems are well known to those of skill in the art [see, eg. Basic Methods in Molecular Biology, L. G. Davis et al, ed., Elsevier Press, NY (1986)]. The labeled EST insert is then used as a probe to screen a lambda phage cDNA library or a plasmid cDNA library, identifying colonies containing clones related to the probe cDNA which can be purified by known methods. The ends of the newly purified clones are then sequenced to identify full length sequences and complete sequencing of full length clones is performed by enzymatic digestion or primer walking. A similar screening and clone selection approach can be applied to clones from a genomic DNA library.

Additionally, an EST or gene identified by this method as associated with inherited disorders can be used to determine at what stage during embryonic development the selected gene from which it is derived is developed by screening embryonic DNA libraries from various stages of development, e.g. 2-cell, 8-cell, etc., for the selected gene. As has been mentioned above, the invention may be applied in addtional temporal modes for monitoring the progression of a disease state, the efficacy of a particular treatment modality or the aging process of an individual.

Thus, the methods of this invention permit the identification, isolation and sequencing of a gene which is differentially expressed in a selected disease/infection. As described in more detail below, the identified gene may then be employed to obtain any protein encoded thereby, or may be employed as a target for diagnostic methods or therapeutic approaches to the treatment of the disease, including, e.g., drug development.

The same methods as described above for the identification of genes, including genes of unknown function, which are differentially expressed in a disease state, may also be employed to identify other genes of interest. For example, another embodiment of this invention includes a method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with that pathogen or the gene of the host which is altered in its expression as a result of the infection.

One such method employs a healthy test surface as described above, employing defined oligonucleotide/polynucleotides from a sample of a healthy, uninfected animal. The second such surface has immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences of ESTs isolated from at least one analogous tissue or cell sample of an infected animal (the "infection test surface"). Polynucleotide sequences are isolated from a biological sample from a healthy animal ("healthy control") and a second sample is similarly

prepared from an animal infected with the selected pathogen ("infection sample"). These two samples are desirably selected from the cell or tissue analogous to that which provided the immobilized oligonucleotide/polynucleotides. It would also be possible to provide samples from the nucleic acid of the pathogen itself.

According to the method the healthy control sample is contacted with one set of the healthy test surface and the infection test surface described above for a time sufficient to permit detectable hybridization to occur between the sample and the immobilized defined oligonucleotide/polynucleotides on each surface. The results of this hybridization are a first hybridization pattern formed between the nucleotides of healthy control and the healthy test surface and a second hybridization pattern formed between the nucleotides of healthy control sample and the infection test surface.

In a similar manner, the infection sample is detectably hybridized to another set of healthy test and infection test surfaces, forming a third hybridization pattern between the infection sample and healthy test surface and a fourth hybridization pattern between the infection sample and the infection test surface.

Comparing the four hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy animal and the animal infected with the pathogen by the presence of differences in the hybridization patterns at pre-defined regions. As mentioned differential expression is not required and simple qualitative analysis is possible by reference to gene expression which is simply present or absent.

A second embodiment of this method parallels the second embodiment of the method as applied to disease above, i.e., the same process is employed, with the exception that plurality of defined oligonucleotide/polynucleotide sequences forming the healthy test sample surface and the infection test sample surface are immobilized on a single solid support. The resulting first hybridization pattern (healthy control sample with healthy/infection test sample) and second hybridization pattern (infection sample with healthy/infection test sample) permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the infection sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained.

As described above for the methods for identifying differential gene expression between diseased and healthy animals, the

oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding ESTs from which the oligonucleotide/polynucleotide sequences are obtained and the genes expressed by the pathogen identified for similar purposes. Other embodiments of these methods may be developed with resort to the teaching herein, by altering the samples which provide the defined oligonucleotide/polynucleotides. For example, an EST, identified with a differentially expressed gene by the method of this invention is also useful in detecting genes expressed in the various stages of an pathogen's development, particularly the infective stage and following the cours of drug treatment and emergence of resistant variants. For example, employing the techniques described above, the EST can be used for detecting a gene in various stages of the parasitic *Plasmodium* species life cycle, which include blood stages, liver stages, and gametocyte stages.

*B. Diagnostic Methods*

In addition to use of the methods and compositions of this invention for identifying differentially expressed genes, another embodiment of this invention provides diagnostic methods for diagnosing a selected disease state, or a selected state resulting from aging, exposure to drugs or infection in an animal. According to this aspect of the invention, a first surface, described as the healthy test surface above, and a second surface, described as the disease test surface or infection test surface, are prepared depending on the disease or infection to be diagnosed. The same processes of detectable hybridization to a first and second set of these surfaces with the healthy control sample and disease/infection sample are followed to provide the four above-described hybridization patterns, i.e., healthy control sample with healthy test surface; healthy control sample with disease/infection test surface; disease/infection sample with healthy test surface; and disease/infection sample with disease/infection test surface.

The diagnosis of disease or infection is provided by comparing the four hybridization patterns. Substantial differences between the first and third hybridization patterns, respectively, and the second and fourth hybridization patterns, respectively, indicate the presence of the selected disease or infection in said animal. Substantial similarities in the first and third hybridization patterns and second and fourth hybridization patterns indicates the absence of disease or infection.

A similar embodiment utilizes the single surface bearing both the healthy test surface defined oligonucleotide/polynucleotides and the disease/infection test surface defined oligonucleotide/polynucleotides as described above. Parallel process steps as described above for detection of genes differentially expressed in disease and infected states are followed, resulting in a first hybridization

pattern (healthy control sample with single healthy and disease/infection test sample) and a second hybridization pattern (disease/infection sample with another copy of the single healthy and disease/infection test sample).

Diagnosis is accomplished by comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicate the presence of the selected disease or infection in the animal being tested. Substantially similar first and second hybridization patterns indicate the absence of disease or infection. This like many of the foregoing embodiments may use known or unknown ESTs derived from many libraries.

*C. Other Methods of the Invention*

As is obvious to one of skill in the art upon reading this disclosure, the compositions and methods of this invention may also be used for other similar purposes. For example, the general methods and compositions may be adapted easily by manipulation of the samples selected to provide the standardized defined oligonucleotide/polynucleotides, and selection of the samples selected for hybridization thereto. One such modification is the use of this invention to identify cell markers of any type, e.g., markers of cancer cells, stem cell markers, and the like. Another modification involves the use of the method and compositions to generate hybridization patterns useful for forensic identification or an 'expression fingerprint' of genes for identification of one member of a species from another. Similarly, the methods of this invention may be adapted for use in tissue matching for transplantation purposes as well as for molecular histology, i.e., to enable diagnosis of disease or disorders in pathology tissue samples such as biopsies. Still another use of this method is in monitoring the effects of development and aging upon the gene expression in a selected animal, by preparing surfaces bearing oligonucleotide/polynucleotides prepared from samples of standardized younger members of the species being tested. Additionally the patient can serve as an internal control by virtue of having the method applied to blood samples every 5-10 years during his lifetime.

Still another intriguing use of this method is in the area of monitoring the effects of drugs on gene expression, both in laboratories and during clinical trials with animal, especially humans. Because the method can be readily adapted by altering the above parameters, it can essentially be employed to identify differentially expressed genes of any organism, at any stage of development, and under the influence of any factor which can affect gene expression.

*IV. The Genes and Proteins Identified*

Application of the compositions and methods of this invention as above described also provide other compositions, such as any isolated gene sequence which is differentially expressed between a normal healthy animal and an animal having a disease or infection. Another embodiment of this invention is any isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal. Similarly an embodiment of this invention is any gene sequence identified by the methods described herein.

These gene sequences may be employed in conventional methods to produce isolated proteins encoded thereby. To produce a protein of this invention, the DNA sequences of a desired gene identified by the use of the methods of this invention or portions thereof are inserted into a suitable expression system. Desirably, a recombinant molecule or vector is constructed in which the polynucleotide sequence encoding the protein is operably linked to a heterologous expression control sequence permitting expression of the human protein. Numerous types of appropriate expression vectors and host cell systems are known in the art for mammalian (including human) expression, insect, e.g., baculovirus expression, yeast, fungal, and bacterial expression, by standard molecular biology techniques.

The transfection of these vectors into appropriate host cells, whether mammalian, bacterial, fungal, or insect, or into appropriate viruses, can result in expression of the selected proteins. Suitable host cells or cell lines for transfection, and viruses, as well as methods for the construction and transfection of such host cells and viruses are well-known. Suitable methods for transfection, culture, amplification, screening, and product production and purification are also known in the art.

The genes and proteins identified by this invention can be employed, if desired in diagnostic compositions useful for the diagnosis of a disease or infection using conventional diagnostic assays. For example, a diagnostic reagent can be developed which detectably targets a gene sequence or protein of this invention in a biological sample of an animal. Such a reagent may be a complementary nucleotide sequence, an antibody (monoclonal, recombinant or polyclonal), or a chemically derived agonist or antagonist. Alternatively, the proteins and polynucleotide sequences of this invention, fragments of same, or complementary sequences thereto, may themselves be useful as diagnostic reagents for diagnosing disease states with which the ESTs of the invention are associated. These reagents may optionally be labelled using diagnostic labels, such as radioactive labels, colorimetric enzyme label systems and the like conventionally used in diagnostic or therapeutic methods, e.g, Northern and Western blotting, antigen-antibody binding and the like. The selection of the appropriate assay format and label system is within the skill of the art and may

readily be chosen without requiring additional explanation by resort to the wealth of art in the diagnostic area.

Additionally, genes and proteins identified according to this invention may be used therapeutically. For example, the EST-containing gene sequences may be useful in gene therapy, to provide a gene sequence which in a disease is not properly or sufficiently expressed. In such a method, a selected gene sequence of this invention is introduced into a suitable vector or other delivery system for delivery to a cell containing a defect in the selected gene. Suitable delivery systems are well known to those of skill in the art and enable the desired EST or gene to be incorporated into the target cell and to be translated by the cell. The EST or gene sequence may be introduced to mutate the existing gene by recombination or provide an active copy thereof in addition to the inactive gene to replace its function.

Alternatively, a protein encoded by an EST or gene of the invention may be useful as a therapeutic reagent for delivery of a biologically active protein, particularly when the disease state is associated with a deficiency of this protein. Such a protein may be incorporated into an appropriate therapeutic formulation, alone or in combination with other active ingredients. Methods of formulating such therapeutic compositions, as well as suitable pharmaceutical carriers, and the like, are well known to those of skill in the art. Still an additional method of delivering the missing protein encoded by an EST, or the gene from which a selected EST was derived, involves expressing it directly *in vivo*. Systems for such *in vivo* expression are well known in the art.

Yet another use of the ESTs, genes identified according to the methods of this invention, or the proteins encoded thereby is a target for the screening and development of natural or synthetic chemical compounds which have utility as therapeutic drugs for the treatment of disease states associated with the identified genes and ESTs derived therefrom. As one example, a compound capable of binding to such a protein encoded by such a gene and either preventing or enhancing its biological activity may be a useful drug component for the treatment or prevention of such disease states.

Conventional assays and techniques may be used for the screening and development of such drugs. As one example, a method for identifying compounds which specifically bind to or inhibit or activate proteins encoded by these gene sequences can include simply the steps of contacting a selected protein or gene product, with a test compound to permit binding of the test compound to the protein; and determining the amount of test compound, if any, which is bound to the protein. Such a method may involve the incubation of the test compound and the protein immobilized on a solid support. Still other conventional methods of drug screening

can involve employing a suitable computer program to determine compounds having similar or complementary chemical structures to that of the gene product or portions thereof and screening those compounds either for competitive binding to the protein to detect enhanced or decreased activity in the presence of the selected compound.

Thus, through use of such methods, the present invention is anticipated to provide compounds capable of interacting with these genes, ESTs, or encoded proteins, or fragments thereof, and either enhancing or decreasing the biological activity, as desired. Such compounds are believed to be encompassed by this invention.

Numerous modifications and variations of the present invention are included in the above-identified specification and are expected to be obvious to one of skill in the art. Such modifications and alterations to the compositions and processes of the present invention are believed to be encompassed in the scope of the claims appended hereto.

WHAT IS CLAIMED IS:

1. A method for identifying genes which are differentially expressed in two different pre-determined states of an organism comprising:

a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample in a first state and present in excess relative to the polynucleotide to be hybridized;

b. providing a second surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample in a second state and present in excess relative to the polynucleotide to be hybridized;

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a said organism in said first state, said sample selected from sources analogous to the sources of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface,

d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from said organism in said second state, said sample selected from sources analogous to the sources of step (c), said hybridization sufficient to form a third and fourth hybridization pattern on each said first and second surface,

e. comparing at least two of the four hybridization patterns, wherein genes differentially expressed in said first and second states are identified by the presence of differences in the hybridization patterns at pre-defined regions;

f. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs or larger gene fragment from which the oligonucleotide/polynucleotides were obtained, whereby identification of the EST or larger gene fragment permits identification of the gene from which the ESTs or larger gene fragment were derived.

2. The method according to Claim 1 wherein said first and second states are respectively healthy and disease; pathogen uninfected and pathogen infected; a first progression state and a second progression of a disease or infection; a first treatment state and a second treatment state of a disease or infection; or a first developmental and a second developmental state.

3. The method according to Claim 1 wherein said organism is a plant or an animal.

4. The method according to Claim 3 wherein said aniaml is a human.

5. A method for identifying genes which are differentially expressed in a normal healthy animal and an animal having a disease comprising:

a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample in a healthy animal and present in excess relative to the polynucleotide to be hybridized;

b. providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample from an animal having said disease and present in excess relative to the polynucleotide to be hybridized;

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from sources analogous to the sources of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface;

d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (c), said hybridization sufficient to form a third and fourth hybridization pattern on each said first and second surface,

e. comparing at least two of the four hybridization patterns, wherein genes differentially expressed in said first and second states are identified by the presence of differences in the hybridization patterns at pre-defined regions;

f. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs or larger gene fragment from which the oligonucleotide/polynucleotides were obtained, whereby identification of the EST or larger gene fragment permits identification of the gene from which the ESTs or larger gene fragment were derived.

6. A method for identifying genes which are differentially expressed in a normal healthy animal and an animal having a disease comprising:

a. providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample in of a healthy animal and an analogous selected sample of an animal having said disease and both present in excess relative to the polynucleotide to be hybridized;

b. detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

c. detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

d. comparing the two hybridization patterns, wherein genes differentially expressed in a disease state are identified by the presence of differences in the hybridization patterns at pre-defined regions;

e. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

7. A method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with said pathogen comprising:

a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample of a healthy, uninfected animal and present in excess relative to the polynucleotide to be hybridized;

b. providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from at least one selected cell, tissue, organ or organism sample of an infected animal;

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form first and second hybridization patterns on each said first and second surface,

d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from an infected animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form third and fourth hybridization patterns on each said first and second surface,

e. comparing the four hybridization patterns, wherein genes of said pathogen which are expressed in an infected animal are identified by the presence of differences in the hybridization patterns at pre-defined regions;

f. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

8. A method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with said pathogen comprising:

a. providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample in of a healthy animal and an analogous selected sample of an animal having said disease and both present in excess relative to the polynucleotide to be hybridized

b. detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

c. detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from a sample from an infected animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

d. comparing the two hybridization patterns, wherein genes of said pathogen which are expressed in an infected animal are identified by the presence of differences in the hybridization patterns at pre-defined regions;

e. identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

9. A composition suitable for use in hybridization comprising a solid surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences for hybridization, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample of a healthy animal, at least one analogous sample of said animal having a disease, at least one analogous sample of said animal infected with a microbial pathogen, and any combination thereof.

10. An isolated gene sequence which is differentially expressed in a normal healthy animal and an animal having a disease, identified by the method of claim 1.

11. An isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal identified by the method of claim 7.

12. A diagnostic composition useful for the diagnosis of a disease comprising a reagent capable of detectably targeting a gene sequence of claim 10 in a biological sample of an animal.

13. A diagnostic composition useful for the diagnosis of infection by a pathogen comprising a reagent capable of detectably targeting a gene sequence of claim 11 in a biological sample of an animal.

14. An isolated protein produced by expression of a gene sequence of claim 10.

15. An isolated pathogen protein produced by expression of a gene sequence of claim 11.

16. A therapeutic composition comprising a protein or fragment thereof selected from the group consisting of a protein of claim 10 and a protein of claim 15.

17. A method for diagnosing a selected disease or infection in an animal comprising:

a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample of a healthy animal and present in excess relative to the polynucleotide to be hybridized;

b. providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence comprising a fragment of an EST isolated from at least one said tissue of an animal having said disease;

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a DNA library prepared from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface;

d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a DNA library prepared from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (c), said hybridization sufficient to form a third and fourth hybridization pattern on each said first and second surface;

e. comparing the four hybridization patterns, wherein substantial differences between the first and third hybridization patterns and the second and fourth hybridization patterns indicates the presence of said selected disease or infection in said animal, and substantial similarities in said first and third hybridization patterns and second and fourth hybridization patterns indicates the absence of disease or infection.

18. A method for diagnosing a selected disease or infection in an animal comprising:

a. providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence comprising a fragment of an EST isolated from a DNA library prepared from the group consisting of a selected cell or tissue sample of a healthy animal and an analogous selected cell or tissue sample of an animal having said disease;

b. detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

c. detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from a DNA library prepared from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

d. comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicates the presence of said selected disease or infection in said animal, and substantial similarities in said first and second hybridization patterns indicates the absence of disease or infection.

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01863

A. CLASSIFICATION OF SUBJECT MATTER
IPC(6) :C12Q 1/68
US CL :435/6
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 435/6

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
APS, CAS, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y | ANALYTICAL BIOCHEMISTRY, VOLUME 187, ISSUED 1990, FARGNOLI ET AL, "LOW-RATIO HYBRIDIZATION SUBTRACTION", PAGES 364-373, SEE ENTIRE DOCUMENT. | 1-18 |
| Y | PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES USA, VOLUME 88, ISSUED MARCH 1991, PATANJALI ET AL, "CONSTRUCTION OF A UNIFORM-ABUNDANCE (NORMALIZED) CDNA LIBRARY", PAGES 1943-1947, SEE ENTIRE DOCUMENT. | 1-18 |
| Y | SCIENCE, VOLUME 245, ISSUED 29 SEPTEMBER 1989, OLSON ET AL. "A COMMON LANGUAGE FOR PHYSICAL MAPPING OF THE HUMAN GENOME", PAGES 1434-1435, SEE ENTIRE DOCUMENT. | 1-18 |

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search
03 APRIL 1995

Date of mailing of the international search report
17MAY1995

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
EGGERTON CAMPBELL
Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01863

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y | SCIENCE, VOLUME 252, ISSUED 21 JUNE 1991, ADAMS ET AL, "COMPLEMENTARY DNA SEQUENCING: EXPRESSED SEQUENCE TAGS AND HUMAN GENOME PROJECT", PAGES 1651-1656, SEE ENTIRE DOCUMENT. | 1-18 |

**PCT** WORLD INTELLECTUAL PROPERTY ORGANIZATI
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT

(51) International Patent Classification 6: C12Q 1/68, G06F 15/00 — A1

(11) International Publication Number: WO 95/20681

(43) International Publication Date: 3 August 1995 (03.08.95)

(21) International Application Number: PCT/US95/01160

(22) International Filing Date: 27 January 1995 (27.01.95)

(30) Priority Data:
08/187,530 27 January 1994 (27.01.94) US
08/282,955 29 July 1994 (29.07.94) US

(71) Applicant: INCYTE PHARMACEUTICALS, INC. [US/US]; 3330 Hillview Avenue, Palo Alto, CA 94304 (US).

(72) Inventors: SEILHAMER, Jeffrey, J.; 12555 La Cresta, Los Altos Hills, CA 94022 (US). SCOTT, Randal, W.; 13140 Sun-Mor, Mountain View, CA 94040 (US).

(74) Agents: CAGE, Kenneth, L. et al.; Willian Brinks Hofer Gilson & Lione, 2000 K Street, N.W., Suite 200, Washington, DC 20006-1809 (US).

(81) Designated States: AM, AU, BB, BG, BR, BY, CA, CN, CZ, EE, FI, GE, HU, JP, KG, KP, KR, KZ, LK, LR, LT, LV, MD, MG, MN, MX, NO, NZ, PL, RO, RU, SI, SK, TJ, TT, UA, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), ARIPO patent (KE, MW, SD, SZ).

**Published**
*With international search report.*

(54) Title: COMPARATIVE GENE TRANSCRIPT ANALYSIS

(57) Abstract

A method and system for quantifying the relative abundance of gene transcripts in a biological specimen. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs (gene transcript imaging analysis). Another embodiment of the method produces a gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, the gene transcript imaging can be used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells. The invention provides a method for comparing the gene transcript image analysis from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens.

*FOR THE PURPOSES OF INFORMATION ONLY*

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

| | | | | | |
|---|---|---|---|---|---|
| **AT** | Austria | **GB** | United Kingdom | **MR** | Mauritania |
| **AU** | Australia | **GE** | Georgia | **MW** | Malawi |
| **BB** | Barbados | **GN** | Guinea | **NE** | Niger |
| **BE** | Belgium | **GR** | Greece | **NL** | Netherlands |
| **BF** | Burkina Faso | **HU** | Hungary | **NO** | Norway |
| **BG** | Bulgaria | **IE** | Ireland | **NZ** | New Zealand |
| **BJ** | Benin | **IT** | Italy | **PL** | Poland |
| **BR** | Brazil | **JP** | Japan | **PT** | Portugal |
| **BY** | Belarus | **KE** | Kenya | **RO** | Romania |
| **CA** | Canada | **KG** | Kyrgystan | **RU** | Russian Federation |
| **CF** | Central African Republic | **KP** | Democratic People's Republic of Korea | **SD** | Sudan |
| **CG** | Congo | | | **SE** | Sweden |
| **CH** | Switzerland | **KR** | Republic of Korea | **SI** | Slovenia |
| **CI** | Côte d'Ivoire | **KZ** | Kazakhstan | **SK** | Slovakia |
| **CM** | Cameroon | **LI** | Liechtenstein | **SN** | Senegal |
| **CN** | China | **LK** | Sri Lanka | **TD** | Chad |
| **CS** | Czechoslovakia | **LU** | Luxembourg | **TG** | Togo |
| **CZ** | Czech Republic | **LV** | Latvia | **TJ** | Tajikistan |
| **DE** | Germany | **MC** | Monaco | **TT** | Trinidad and Tobago |
| **DK** | Denmark | **MD** | Republic of Moldova | **UA** | Ukraine |
| **ES** | Spain | **MG** | Madagascar | **US** | United States of America |
| **FI** | Finland | **ML** | Mali | **UZ** | Uzbekistan |
| **FR** | France | **MN** | Mongolia | **VN** | Viet Nam |
| **GA** | Gabon | | | | |

# COMPARATIVE GENE TRANSCRIPT ANALYSIS

## 1. FIELD OF INVENTION

The present invention is in the field of molecular biology and computer science; more particularly, the present invention describes methods of analyzing gene transcripts and diagnosing the genetic expression of cells and tissue.

## 2. BACKGROUND OF THE INVENTION

Until very recently, the history of molecular biology has been written one gene at a time. Scientists have observed the cell's physical changes, isolated mixtures from the cell or its milieu, purified proteins, sequenced proteins and therefrom constructed probes to look for the corresponding gene.

Recently, different nations have set up massive projects to sequence the billions of bases in the human genome. These projects typically begin with dividing the genome into large portions of chromosomes and then determining the sequences of these pieces, which are then analyzed for identity with known proteins or portions thereof, known as motifs. Unfortunately, the majority of genomic DNA does not encode proteins and though it is postulated to have some effect on the cell's ability to make protein, its relevance to medical applications is not understood at this time.

A third methodology involves sequencing only the transcripts encoding the cellular machinery actively involved in making protein, namely the mRNA. The advantage is that the cell has already edited out all the non-coding DNA, and it is relatively easy to identify the protein-coding portion of the RNA. The utility of this approach was not immediately obvious to genomic researchers. In fact, when cDNA sequencing was initially proposed, the method was roundly denounced by those committed to genomic sequencing. For example, the head of the U.S. Human Genome project discounted CDNA sequencing as not valuable and refused to approve funding of projects.

In this disclosure, we teach methods for analyzing DNA, including cDNA libraries. Based on our analyses and

research, we see each individual gene product as a "pixel" of information, which relates to the expression of that, and only that, gene. We teach herein, methods whereby the individual "pixels" of gene expression information can be combined into a single gene transcript "image," in which each of the individual genes can be visualized simultaneously and allowing relationships between the gene pixels to be easily visualized and understood.

We further teach a new method which we call electronic subtraction. Electronic subtraction will enable the gene researcher to turn a single image into a moving picture, one which describes the temporality or dynamics of gene expression, at the level of a cell or a whole tissue. It is that sense of "motion" of cellular machinery on the scale of a cell or organ which constitutes the new invention herein. This constitutes a new view into the process of living cell physiology and one which holds great promise to unveil and discover new therapeutic and diagnostic approaches in medicine.

We teach another method which we call "electronic northern," which tracks the expression of a single gene across many types of cells and tissues.

Nucleic acids (DNA and RNA) carry within their sequence the hereditary information and are therefore the prime molecules of life. Nucleic acids are found in all living organisms including bacteria, fungi, viruses, plants and animals. It is of interest to determine the relative abundance of different discrete nucleic acids in different cells, tissues and organisms over time under various conditions, treatments and regimes.

All dividing cells in the human body contain the same set of 23 pairs of chromosomes. It is estimated that these autosomal and sex chromosomes encode approximately 100,000 genes. The differences among different types of cells are believed to reflect the differential expression of the 100,000 or so genes. Fundamental questions of biology could be answered by understanding which genes are transcribed and knowing the relative abundance of transcripts in different cells.

Previously, the art has only provided for the analysis of a few known genes at a time by standard molecular biology techniques such as PCR, northern blot analysis, or other types of DNA probe analysis such as in situ hybridization. Each of these methods allows one to analyze the transcription of only known genes and/or small numbers of genes at a time. Nucl. Acids Res. 19, 7097-7104 (1991); Nucl. Acids Res. 18, 4833-42 (1990); Nucl. Acids Res. 18, 2789-92 (1989); European J. Neuroscience 2, 1063-1073 (1990); Analytical Biochem. 187, 364-73 (1990); Genet. Annals Techn. Appl. 7, 64-70 (1990); GATA 8(4), 129-33 (1991); Proc. Natl. Acad. Sci. USA 85, 1696-1700 (1988); Nucl. Acids Res. 19, 1954 (1991); Proc. Natl. Acad. Sci. USA 88, 1943-47 (1991); Nucl. Acids Res. 19, 6123-27 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42 (1988); Nucl. Acids Res. 16, 10937 (1988).

Studies of the number and types of genes whose transcription is induced or otherwise regulated during cell processes such as activation, differentiation, aging, viral transformation, morphogenesis, and mitosis have been pursued for many years, using a variety of methodologies. One of the earliest methods was to isolate and analyze levels of the proteins in a cell, tissue, organ system, or even organisms both before and after the process of interest. One method of analyzing multiple proteins in a sample is using 2-dimensional gel electrophoresis, wherein proteins can be, in principle, identified and quantified as individual bands, and ultimately reduced to a discrete signal. At present, 2-dimensional analysis only resolves approximately 15% of the proteins. In order to positively analyze those bands which are resolved, each band must be excised from the membrane and subjected to protein sequence analysis using Edman degradation. Unfortunately, most of the bands were present in quantities too small to obtain a reliable sequence, and many of those bands contained more than one discrete protein. An additional difficulty is that many of the proteins were blocked at the amino-terminus, further complicating the sequencing process.

Analyzing differentiation at the gene transcription level has overcome many of these disadvantages and drawbacks, since the power of recombinant DNA technology allows amplification of signals containing very small amounts of material. The most common method, called "hybridization subtraction," involves isolation of mRNA from the biological specimen before (B) and after (A) the developmental process of interest, transcribing one set of mRNA into cDNA, subtracting specimen B from specimen A (mRNA from cDNA) by hybridization, and constructing a cDNA library from the non-hybridizing mRNA fraction. Many different groups have used this strategy successfully, and a variety of procedures have been published and improved upon using this same basic scheme. Nucl. Acids Res. 19, 7097-7104 (1991); Nucl. Acids Res. 18, 4833-42 (1990); Nucl. Acids Res. 18, 2789-92 (1989); European J. Neuroscience 2, 1063-1073 (1990); Analytical Biochem. 187, 364-73 (1990); Genet. Annals Techn. Appl. 7, 64-70 (1990); GATA 8(4), 129-33 (1991); Proc. Natl. Acad. Sci. USA 85, 1696-1700 (1988); Nucl. Acids Res. 19, 1954 (1991); Proc. Natl. Acad. Sci. USA 88, 1943-47 (1991); Nucl. Acids Res. 19, 6123-27 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42 (1988); Nucl. Acids Res. 16, 10937 (1988).

Although each of these techniques have particular strengths and weaknesses, there are still some limitations and undesirable aspects of these methods: First, the time and effort required to construct such libraries is quite large. Typically, a trained molecular biologist might expect construction and characterization of such a library to require 3 to 6 months, depending on the level of skill, experience, and luck. Second, the resulting subtraction libraries are typically inferior to the libraries constructed by standard methodology. A typical conventional cDNA library should have a clone complexity of at least $10^6$ clones, and an average insert size of 1-3 kB. In contrast, subtracted libraries can have complexities of $10^2$ or $10^3$ and average insert sizes of 0.2 kB. Therefore, there can be a significant loss of clone and sequence information associated with such libraries. Third, this

approach allows the researcher to capture only the genes induced in specimen A relative to specimen B, not vice-versa, nor does it easily allow comparison to a third specimen of interest (C). Fourth, this approach requires very large amounts (hundreds of micrograms) of "driver" mRNA (specimen B), which significantly limits the number and type of subtractions that are possible since many tissues and cells are very difficult to obtain in large quantities.

Fifth, the resolution of the subtraction is dependent upon the physical properties of DNA:DNA or RNA:DNA hybridization. The ability of a given sequence to find a hybridization match is dependent on its unique CoT value. The CoT value is a function of the number of copies (concentration) of the particular sequence, multiplied by the time of hybridization. It follows that for sequences which are abundant, hybridization events will occur very rapidly (low CoT value), while rare sequences will form duplexes at very high CoT values. CoT values which allow such rare sequences to form duplexes and therefore be effectively selected are difficult to achieve in a convenient time frame. Therefore, hybridization subtraction is simply not a useful technique with which to study relative levels of rare mRNA species. Sixth, this problem is further complicated by the fact that duplex formation is also dependent on the nucleotide base composition for a given sequence. Those sequences rich in G + C form stronger duplexes than those with high contents of A + T. Therefore, the former sequences will tend to be removed selectively by hybridization subtraction. Seventh, it is possible that hybridization between nonexact matches can occur. When this happens, the expression of a homologous gene may "mask" expression of a gene of interest, artificially skewing the results for that particular gene.

Matsubara and Okubo proposed using partial cDNA sequences to establish expression profiles of genes which could be used in functional analyses of the human genome. Matsubara and Okubo warned against using random priming, as

it creates multiple unique DNA fragments from individual mRNAs and may thus skew the analysis of the number of particular mRNAs per library. They sequenced randomly selected members from a 3'-directed cDNA library and established the frequency of appearance of the various ESTs. They proposed comparing lists of ESTs from various cell types to classify genes. Genes expressed in many different cell types were labeled housekeepers and those selectively expressed in certain cells were labeled cell-specific genes, even in the absence of the full sequence of the gene or the biological activity of the gene product.

The present invention avoids the drawbacks of the prior art by providing a method to quantify the relative abundance of multiple gene transcripts in a given biological specimen by the use of high-throughput sequence-specific analysis of individual RNAs and/or their corresponding cDNAs.

The present invention offers several advantages over current protein discovery methods which attempt to isolate individual proteins based upon biological effects. The method of the instant invention provides for detailed diagnostic comparisons of cell profiles revealing numerous changes in the expression of individual transcripts.

The instant invention provides several advantages over current subtraction methods including a more complex library analysis ($10^6$ to $10^7$ clones as compared to $10^3$ clones) which allows identification of low abundance messages as well as enabling the identification of messages which either increase or decrease in abundance. These large libraries are very routine to make in contrast to the libraries of previous methods. In addition, homologues can easily be distinguished with the method of the instant invention.

This method is very convenient because it organizes a large quantity of data into a comprehensible, digestible format. The most significant differences are highlighted by electronic subtraction. In depth analyses are made more convenient.

The present invention provides several advantages over previous methods of electronic analysis of cDNA. The method is particularly powerful when more than 100 and preferably more than 1,000 gene transcripts are analyzed. In such a case, new low-frequency transcripts are discovered and tissue typed.

High resolution analysis of gene expression can be used directly as a diagnostic profile or to identify disease-specific genes for the development of more classic diagnostic approaches.

This process is defined as gene transcript frequency analysis. The resulting quantitative analysis of the gene transcripts is defined as comparative gene transcript analysis.

## 3. SUMMARY OF THE INVENTION

The invention is a method of analyzing a specimen containing gene transcripts comprising the steps of (a) producing a library of biological sequences; (b) generating a set of transcript sequences, where each of the transcript sequences in said set is indicative of a different one of the biological sequences of the library; (c) processing the transcript sequences in a programmed computer (in which a database of reference transcript sequences indicative of reference sequences is stored), to generate an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of sequence annotation and a degree of match between one of the biological sequences of the library and at least one of the reference sequences; and (d) processing each said identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

The invention also includes a method of comparing two specimens containing gene transcripts. The first specimen is processed as described above. The second specimen is used to produce a second library of biological sequences, which is used to generate a second set of transcript sequences, where each of the transcript sequences in the

second set is indicative of one of the biological sequences of the second library. Then the second set of transcript sequences is processed in a programmed computer to generate a second set of identified sequence values, namely the further identified sequence values, each of which is indicative of a sequence annotation and includes a degree of match between one of the biological sequences of the second library and at least one of the reference sequences. The further identified sequence values are processed to generate further final data values indicative of the number of times each further identified sequence value is present in the second library. The final data values from the first specimen and the further identified sequence values from the second specimen are processed to generate ratios of transcript sequences, which indicate the differences in the number of gene transcripts between the two specimens.

In a further embodiment, the method includes quantifying the relative abundance of mRNA in a biological specimen by (a) isolating a population of mRNA transcripts from a biological specimen; (b) identifying genes from which the mRNA was transcribed by a sequence-specific method; (c) determining the numbers of mRNA transcripts corresponding to each of the genes; and (d) using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts.

Also disclosed is a method of producing a gene transcript image analysis by first obtaining a mixture of mRNA, from which cDNA copies are made. The cDNA is inserted into a suitable vector which is used to transfect suitable host strain cells which are plated out and permitted to grow into clones, each cone representing a unique mRNA. A representative population of clones transfected with cDNA is isolated. Each clone in the population is identified by a sequence-specific method which identifies the gene from which the unique mRNA was transcribed. The number of times each gene is identified to a clone is determined to evaluate gene transcript abundance. The genes and their abundances are listed in order of abundance to produce a gene transcript image.

In a further embodiment, the relative abundance of the gene transcripts in one cell type or tissue is compared with the relative abundance of gene transcript numbers in a second cell type or tissue in order to identify the differences and similarities.

In a further embodiment, the method includes a system for analyzing a library of biological sequences including a means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a different one of the biological sequences of the library; and a means for processing the transcript sequences in a computer system in which a database of reference transcript sequences indicative of reference sequences is stored, wherein the computer is programmed with software for generating an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and the degree of match between a different one of the biological sequences of the library and at least one of the reference sequences, and for processing each said identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

In essence, the invention is a method and system for quantifying the relative abundance of gene transcripts in a biological specimen. The invention provides a method for comparing the gene transcript image from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens. Thus, this gene transcript image and its comparison can be used as a diagnostic. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs: a gene transcript image. Another embodiment of the method produces the gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, two or more gene transcript images can be compared and used to detect or diagnose a particular biological state, disease,

or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells.

## 4. DESCRIPTION OF THE TABLES AND DRAWINGS

### 4.1. TABLES

Table 1 presents a detailed explanation of the letter codes utilized in Tables 2-5.

Table 2 lists the one hundred most common gene transcripts. It is a partial list of isolates from the HUVEC cDNA library prepared and sequenced as described below. The left-hand column refers to the sequence's order of abundance in this table. The next column labeled "number" is the clone number of the first HUVEC sequence identification reference matching the sequence in the "entry" column number. Isolates that have not been sequenced are not present in Table 2. The next column, labeled "N", indicates the total number of cDNAs which have the same degree of match with the sequence of the reference transcript in the "entry" column.

The column labeled "entry" gives the NIH GENBANK locus name, which corresponds to the library sequence numbers. The "s" column indicates in a few cases the species of the reference sequence. The code for column "s" is given in Table 1. The column labeled "descriptor" provides a plain English explanation of the identity of the sequence corresponding to the NIH GENBANK locus name in the "entry" column.

Table 3 is a comparison of the top fifteen most abundant gene transcripts in normal monocytes and activated macrophage cells.

Table 4 is a detailed summary of library subtraction analysis summary comparing the THP-1 and human macrophage cDNA sequences. In Table 4, the same code as in Table 2 is used. Additional columns are for "bgfreq" (abundance number in the subtractant library), "rfend" (abundance number in the target library) and "ratio" (the target abundance number divided by the subtractant abundance number). As is clear from perusal of the table, when the abundance number in the subtractant library is "0", the

target abundance number is divided by 0.05. This is a way of obtaining a result (not possible dividing by 0) and distinguishing the result from ratios of subtractant numbers of 1.

Table 5 is the computer program, written in source code, for generating gene transcript subtraction profiles.

Table 6 is a partial listing of database entries used in the electronic northern blot analysis as provided by the present invention.

## 4.2. BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a chart summarizing data collected and stored regarding the library construction portion of sequence preparation and analysis.

Figure 2 is a diagram representing the sequence of operations performed by "abundance sort" software in a class of preferred embodiments of the inventive method.

Figure 3 is a block diagram of a preferred embodiment of the system of the invention.

Figure 4 is a more detailed block diagram of the bioinformatics process from new sequence (that has already been sequenced but not identified) to printout of the transcript imaging analysis and the provision of database subscriptions.

## 5. DETAILED DESCRIPTION OF THE INVENTION

The present invention provides a method to compare the relative abundance of gene transcripts in different biological specimens by the use of high-throughput sequence-specific analysis of individual RNAs or their corresponding cDNAs (or alternatively, of data representing other biological sequences). This process is denoted herein as gene transcript imaging. The quantitative analysis of the relative abundance for a set of gene transcripts is denoted herein as "gene transcript image analysis" or "gene transcript frequency analysis". The present invention allows one to obtain a profile for gene transcription in any given population of cells or tissue from any type of organism. The invention can be applied to

obtain a profile of a specimen consisting of a single cell (or clones of a single cell), or of many cells, or of tissue more complex than a single cell and containing multiple cell types, such as liver.

The invention has significant advantages in the fields of diagnostics, toxicology and pharmacology, to name a few. A highly sophisticated diagnostic test can be performed on the ill patient in whom a diagnosis has not been made. A biological specimen consisting of the patient's fluids or tissues is obtained, and the gene transcripts are isolated and expanded to the extent necessary to determine their identity. Optionally, the gene transcripts can be converted to cDNA. A sampling of the gene transcripts are subjected to sequence-specific analysis and quantified. These gene transcript sequence abundances are compared against reference database sequence abundances including normal data sets for diseased and healthy patients. The patient has the disease(s) with which the patient's data set most closely correlates.

For example, gene transcript frequency analysis can be used to differentiate normal cells or tissues from diseased cells or tissues, just as it highlights differences between normal monocytes and activated macrophages in Table 3.

In toxicology, a fundamental question is which tests are most effective in predicting or detecting a toxic effect. Gene transcript imaging provides highly detailed information on the cell and tissue environment, some of which would not be obvious in conventional, less detailed screening methods. The gene transcript image is a more powerful method to predict drug toxicity and efficacy. Similar benefits accrue in the use of this tool in pharmacology. The gene transcript image can be used selectively to look at protein categories which are expected to be affected, for example, enzymes which detoxify toxins.

In an alternative embodiment, comparative gene transcript frequency analysis is used to differentiate between cancer cells which respond to anti-cancer agents and those which do not respond. Examples of anti-cancer

agents are tamoxifen, vincristine, vinblastine, podophyllotoxins, etoposide, tenisposide, cisplatin, biologic response modifiers such as interferon, Il-2, GM-CSF, enzymes, hormones and the like. This method also provides a means for sorting the gene transcripts by functional category. In the case of cancer cells, transcription factors or other essential regulatory molecules are very important categories to analyze across different libraries.

In yet another embodiment, comparative gene transcript frequency analysis is used to differentiate between control liver cells and liver cells isolated from patients treated with experimental drugs like FIAU to distinguish between pathology caused by the underlying disease and that caused by the drug.

In yet another embodiment, comparative gene transcript frequency analysis is used to differentiate between brain tissue from patients treated and untreated with lithium.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between cyclosporin and FK506-treated cells and normal cells.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between virally infected (including HIV-infected) human cells and uninfected human cells. Gene transcript frequency analysis is also used to rapidly survey gene transcripts in HIV-resistant, HIV-infected, and HIV-sensitive cells. Comparison of gene transcript abundance will indicate the success of treatment and/or new avenues to study.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between bronchial lavage fluids from healthy and unhealthy patients with a variety of ailments.

In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between cell, plant, microbial and animal mutants and wild-type species. In addition, the transcript abundance program is adapted to permit the scientist to evaluate the transcription of one gene in many different tissues. Such comparisons could

identify deletion mutants which do not produce a gene product and point mutants which produce a less abundant or otherwise different message. Such mutations can affect basic biochemical and pharmacological processes, such as mineral nutrition and metabolism, and can be isolated by means known to those skilled in the art. Thus, crops with improved yields, pest resistance and other factors can be developed.

In a further embodiment, comparative gene transcript frequency analysis is used for an interspecies comparative analysis which would allow for the selection of better pharmacologic animal models. In this embodiment, humans and other animals (such as a mouse), or their cultured cells are treated with a specific test agent. The relative sequence abundance of each cDNA population is determined. If the animal test system is a good model, homologous genes in the animal cDNA population should change expression similarly to those in human cells. If side effects are detected with the drug, a detailed transcript abundance analysis will be performed to survey gene transcript changes. Models will then be evaluated by comparing basic physiological changes.

In a further embodiment, comparative gene transcript frequency analysis is used in a clinical setting to give a highly detailed gene transcript profile of a patient's cells or tissue (for example, a blood sample). In particular, gene transcript frequency analysis is used to give a high resolution gene expression profile of a diseased state or condition.

In the preferred embodiment, the method utilizes high-throughput cDNA sequencing to identify specific transcripts of interest. The generated cDNA and deduced amino acid sequences are then extensively compared with GENBANK and other sequence data banks as described below. The method offers several advantages over current protein discovery by two-dimensional gel methods which try to identify individual proteins involved in a particular biological effect. Here, detailed comparisons of profiles of activated and inactive cells reveal numerous changes in

the expression of individual transcripts. After it is determined if the sequence is an "exact" match, similar or a non-match, the sequence is entered into a database. Next, the numbers of copies of cDNA corresponding to each gene are tabulated. Although this can be done slowly and arduously, if at all, by human hand from a printout of all entries, a computer program is a useful and rapid way to tabulate this information. The numbers of cDNA copies (optionally divided by the total number of sequences in the data set) provides a picture of the relative abundance of transcripts for each corresponding gene. The list of represented genes can then be sorted by abundance in the cDNA population. A multitude of additional types of comparisons or dimensions are possible and are exemplified below.

An alternate method of producing a gene transcript image includes the steps of obtaining a mixture of test mRNA and providing a representative array of unique probes whose sequences are complementary to at least some of the test mRNAs. Next, a fixed amount of the test mRNA is added to the arrayed probes. The test mRNA is incubated with the probes for a sufficient time to allow hybrids of the test mRNA and probes to form. The mRNA-probe hybrids are detected and the quantity determined. The hybrids are identified by their location in the probe array. The quantity of each hybrid is summed to give a population number. Each hybrid quantity is divided by the population number to provide a set of relative abundance data termed a gene transcript image analysis.

## 6. EXAMPLES

The examples below are provided to illustrate the subject invention. These examples are provided by way of illustration and are not included for the purpose of limiting the invention.

### 6.1. TISSUE SOURCES AND CELL LINES

For analysis with the computer program claimed herein, biological sequences can be obtained from virtually any

source. Most popular are tissues obtained from the human body. Tissues can be obtained from any organ of the body, any age donor, any abnormality or any immortalized cell line. Immortal cell lines may be preferred in some instances because of their purity of cell type; other tissue samples invariably include mixed cell types. A special technique is available to take a single cell (for example, a brain cell) and harness the cellular machinery to grow up sufficient cDNA for sequencing by the techniques and analysis described herein (cf. U.S. Patent Nos. 5,021,335 and 5,168,038, which are incorporated by reference). The examples given herein utilized the following immortalized cell lines: monocyte-like U-937 cells, activated macrophage-like THP-1 cells, induced vascular endothelial cells (HUVEC cells) and mast cell-like HMC-1 cells.

The U-937 cell line is a human histiocytic lymphoma cell line with monocyte characteristics, established from malignant cells obtained from the pleural effusion of a patient with diffuse histiocytic lymphoma (Sundstrom, C. and Nilsson, K. (1976) Int. J. Cancer 17:565). U-937 is one of only a few human cell lines with the morphology, cytochemistry, surface receptors and monocyte-like characteristics of histiocytic cells. These cells can be induced to terminal monocytic differentiation and will express new cell surface molecules when activated with supernatants from human mixed lymphocyte cultures. Upon this type of in vitro activation, the cells undergo morphological and functional changes, including augmentation of antibody-dependent cellular cytotoxicity (ADCC) against erythroid and tumor target cells (one of the principal functions of macrophages). Activation of U-937 cells with phorbol 12-myristate 13-acetate (PMA) in vitro stimulates the production of several compounds, including prostaglandins, leukotrienes and platelet-activating factor (PAF), which are potent inflammatory mediators. Thus, U-937 is a cell line that is well suited for the identification and isolation of gene transcripts associated with normal monocytes.

The HUVEC cell line is a normal, homogeneous, well characterized, early passage endothelial cell culture from human umbilical vein (Cell Systems Corp., 12815 NE 124th Street, Kirkland, WA 98034). Only gene transcripts from induced, or treated, HUVEC cells were sequenced. One batch of 1 X $10^8$ cells was treated for 5 hours with 1 U/ml rIL-1b and 100 ng/ml E.coli lipopolysaccharide (LPS) endotoxin prior to harvesting. A separate batch of 2 X $10^8$ cells was treated at confluence with 4 U/ml TNF and 2 U/ml interferon-gamma (IFN-gamma) prior to harvesting.

THP-1 is a human leukemic cell line with distinct monocytic characteristics. This cell line was derived from the blood of a 1-year-old boy with acute monocytic leukemia (Tsuchiya, S. et al. (1980) Int. J. Cancer: 171-76). The following cytological and cytochemical criteria were used to determine the monocytic nature of the cell line: 1) the presence of alpha-naphthyl butyrate esterase activity which could be inhibited by sodium fluoride; 2) the production of lysozyme; 3) the phagocytosis of latex particles and sensitized SRBC (sheep red blood cells); and 4) the ability of mitomycin C-treated THP-1 cells to activate T-lymphocytes following ConA (concanavalin A) treatment. Morphologically, the cytoplasm contained small azurophilic granules and the nucleus was indented and irregularly shaped with deep folds. The cell line had Fc and C3b receptors, probably functioning in phagocytosis. THP-1 cells treated with the tumor promoter 12-o-tetradecanoyl-phorbol-13 acetate (TPA) stop proliferating and differentiate into macrophage-like cells which mimic native monocyte-derived macrophages in several respects. Morphologically, as the cells change shape, the nucleus becomes more irregular and additional phagocytic vacuoles appear in the cytoplasm. The differentiated THP-1 cells also exhibit an increased adherence to tissue culture plastic.

HMC-1 cells (a human mast cell line) were established from the peripheral blood of a Mayo Clinic patient with mast cell leukemia (Leukemia Res. (1988) 12:345-55). The cultured cells looked similar to immature cloned murine

mast cells, contained histamine, and stained positively for chloroacetate esterase, amino caproate esterase, eosinophil major basic protein (MBP) and tryptase. The HMC-1 cells have, however, lost the ability to synthesize normal IgE receptors. HMC-1 cells also possess a 10;16 translocation, present in cells initially collected by leukophoresis from the patient and not an artifact of culturing. Thus, HMC-1 cells are a good model for mast cells.

## 6.2. CONSTRUCTION OF cDNA LIBRARIES

For inter-library comparisons, the libraries must be prepared in similar manners. Certain parameters appear to be particularly important to control. One such parameter is the method of isolating mRNA. It is important to use the same conditions to remove DNA and heterogeneous nuclear RNA from comparison libraries. Size fractionation of cDNA must be carefully controlled. The same vector preferably should be used for preparing libraries to be compared. At the very least, the same type of vector (e.g., unidirectional vector) should be used to assure a valid comparison. A unidirectional vector may be preferred in order to more easily analyze the output.

It is preferred to prime only with oligo dT unidirectional primer in order to obtain one only clone per mRNA transcript when obtaining cDNAs. However, it is recognized that employing a mixture of oligo dT and random primers can also be advantageous because such a mixture results in more sequence diversity when gene discovery also is a goal. Similar effects can be obtained with DR2 (Clontech) and HXLOX (US Biochemical) and also vectors from Invitrogen and Novagen. These vectors have two requirements. First, there must be primer sites for commercially available primers such as T3 or M13 reverse primers. Second, the vector must accept inserts up to 10 kB.

It also is important that the clones be randomly sampled, and that a significant population of clones is used. Data have been generated with 5,000 clones; however, if very rare genes are to be obtained and/or their relative

abundance determined, as many as 100,000 clones from a single library may need to be sampled. Size fractionation of cDNA also must be carefully controlled. Alternately, plaques can be selected, rather than clones.

Besides the Uni-ZAP™ vector system by Stratagene disclosed below, it is now believed that other similarly unidirectional vectors also can be used. For example, it is believed that such vectors include but are not limited to DR2 (Clontech), and HXLOX (U.S. Biochemical).

Preferably, the details of library construction (as shown in Figure 1) are collected and stored in a database for later retrieval relative to the sequences being compared. Fig. 1 shows important information regarding the library collaborator or cell or cDNA supplier, pretreatment, biological source, culture, mRNA preparation and cDNA construction. Similarly detailed information about the other steps is beneficial in analyzing sequences and libraries in depth.

RNA must be harvested from cells and tissue samples and cDNA libraries are subsequently constructed. cDNA libraries can be constructed according to techniques known in the art. (See, for example, Maniatis, T. et al. (1982) Molecular Cloning, Cold Spring Harbor Laboratory, New York). cDNA libraries may also be purchased. The U-937 cDNA library (catalog No. 937207) was obtained from Stratagene, Inc., 11099 M. Torrey Pines Rd., La Jolla, CA 92037.

The THP-1 cDNA library was custom constructed by Stratagene from THP-1 cells cultured 48 hours with 100 nm TPA and 4 hours with 1 μg/ml LPS. The human mast cell HMC-1 cDNA library was also custom constructed by Stratagene from cultured HMC-1 cells. The HUVEC cDNA library was custom constructed by Stratagene from two batches of induced HUVEC cells which were separately processed.

Essentially, all the libraries were prepared in the same manner. First, poly(A+)RNA (mRNA) was purified. For the U-937 and HMC-1 RNA, cDNA synthesis was only primed with oligo dT. For the THP-1 and HUVEC RNA, cDNA synthesis was primed separately with both oligo dT and random

hexamers, and the two cDNA libraries were treated separately. Synthetic adaptor oligonucleotides were ligated onto cDNA ends enabling its insertion into the Uni-Zap™ vector system (Stratagene), allowing high efficiency unidirectional (sense orientation) lambda library construction and the convenience of a plasmid system with blue-white color selection to detect clones with cDNA insertions. Finally, the two libraries were combined into a single library by mixing equal numbers of bacteriophage.

The libraries can be screened with either DNA probes or antibody probes and the pBluescript® phagemid (Stratagene) can be rapidly excised in vivo. The phagemid allows the use of a plasmid system for easy insert characterization, sequencing, site-directed mutagenesis, the creation of unidirectional deletions and expression of fusion proteins. The custom-constructed library phage particles were infected into E. coli host strain XL1-Blue® (Stratagene), which has a high transformation efficiency, increasing the probability of obtaining rare, under-represented clones in the cDNA library.

### 6.3. ISOLATION OF cDNA CLONES

The phagemid forms of individual cDNA clones were obtained by the in vivo excision process, in which the host bacterial strain was coinfected with both the lambda library phage and an f1 helper phage. Proteins derived from both the library-containing phage and the helper phage nicked the lambda DNA, initiated new DNA synthesis from defined sequences on the lambda target DNA and created a smaller, single stranded circular phagemid DNA molecule that included all DNA sequences of the pBluescript® plasmid and the cDNA insert. The phagemid DNA was secreted from the cells and purified, then used to re-infect fresh host cells, where the double stranded phagemid DNA was produced. Because the phagemid carries the gene for beta-lactamase, the newly-transformed bacteria are selected on medium containing ampicillin.

Phagemid DNA was purified using the Magic Minipreps™ DNA Purification System (Promega catalogue #A7100. Promega

Corp., 2800 Woods Hollow Rd., Madison, WI 53711). This small-scale process provides a simple and reliable method for lysing the bacterial cells and rapidly isolating purified phagemid DNA using a proprietary DNA-binding resin. The DNA was eluted from the purification resin already prepared for DNA sequencing and other analytical manipulations.

Phagemid DNA was also purified using the QIAwell-8 Plasmid Purification System from QIAGEN® DNA Purification System (QIAGEN Inc., 9259 Eton Ave., Chattsworth, CA 91311). This product line provides a convenient, rapid and reliable high-throughput method for lysing the bacterial cells and isolating highly purified phagemid DNA using QIAGEN anion-exchange resin particles with EMPORE™ membrane technology from 3M in a multiwell format. The DNA was eluted from the purification resin already prepared for DNA sequencing and other analytical manipulations.

An alternate method of purifying phagemid has recently become available. It utilizes the Miniprep Kit (Catalog No. 77468, available from Advanced Genetic Technologies Corp., 19212 Orbit Drive, Gaithersburg, Maryland). This kit is in the 96-well format and provides enough reagents for 960 purifications. Each kit is provided with a recommended protocol, which has been employed except for the following changes. First, the 96 wells are each filled with only 1 ml of sterile terrific broth with carbenicillin at 25 mg/L and glycerol at 0.4%. After the wells are inoculated, the bacteria are cultured for 24 hours and lysed with 60 µl of lysis buffer. A centrifugation step (2900 rpm for 5 minutes) is performed before the contents of the block are added to the primary filter plate. The optional step of adding isopropanol to TRIS buffer is not routinely performed. After the last step in the protocol, samples are transferred to a Beckman 96-well block for storage.

Another new DNA purification system is the WIZARD™ product line which is available from Promega (catalog No. A7071) and may be adaptable to the 96-well format.

## 6.4. SEQUENCING OF cDNA CLONES

The cDNA inserts from random isolates of the U-937 and THP-1 libraries were sequenced in part. Methods for DNA sequencing are well known in the art. Conventional enzymatic methods employ DNA polymerase Klenow fragment, Sequenase™ or Taq polymerase to extend DNA chains from an oligonucleotide primer annealed to the DNA template of interest. Methods have been developed for the use of both single- and double-stranded templates. The chain termination reaction products are usually electrophoresed on urea-acrylamide gels and are detected either by autoradiography (for radionuclide-labeled precursors) or by fluorescence (for fluorescent-labeled precursors). Recent improvements in mechanized reaction preparation, sequencing and analysis using the fluorescent detection method have permitted expansion in the number of sequences that can be determined per day (such as the Applied Biosystems 373 and 377 DNA sequencer, Catalyst 800). Currently with the system as described, read lengths range from 250 to 400 bases and are clone dependent. Read length also varies with the length of time the gel is run. In general, the shorter runs tend to truncate the sequence. A minimum of only about 25 to 50 bases is necessary to establish the identification and degree of homology of the sequence. Gene transcript imaging can be used with any sequence-specific method, including, but not limited to hybridization, mass spectroscopy, capillary electrophoresis and 505 gel electrophoresis.

## 6.5. HOMOLOGY SEARCHING OF cDNA CLONE AND DEDUCED PROTEIN (and Subsequent Steps)

Using the nucleotide sequences derived from the cDNA clones as query sequences (sequences of a Sequence Listing), databases containing previously identified sequences are searched for areas of homology (similarity). Examples of such databases include Genbank and EMBL. We next describe examples of two homology search algorithms that can be used, and then describe the subsequent computer-implemented steps to be performed in accordance with preferred embodiments of the invention.

In the following description of the computer-implemented steps of the invention, the word "library" denotes a set (or population) of biological specimen nucleic acid sequences. A "library" can consist of cDNA sequences, RNA sequences, or the like, which characterize a biological specimen. The biological specimen can consist of cells of a single human cell type (or can be any of the other above-mentioned types of specimens). We contemplate that the sequences in a library have been determined so as to accurately represent or characterize a biological specimen (for example, they can consist of representative cDNA sequences from clones of RNA taken from a single human cell).

In the following description of the computer-implemented steps of the invention, the expression "database" denotes a set of stored data which represent a collection of sequences, which in turn represent a collection of biological reference materials. For example, a database can consist of data representing many stored cDNA sequences which are in turn representative of human cells infected with various viruses, cells of humans of various ages, cells from different mammalian species, and so on.

In preferred embodiments, the invention employs a computer programmed with software (to be described) for performing the following steps:

(a) processing data indicative of a library of cDNA sequences (generated as a result of high-throughput cDNA sequencing or other method) to determine whether each sequence in the library matches a DNA sequence of a reference database of DNA sequences (and if so, identifying the reference database entry which matches the sequence and indicating the degree of match between the reference sequence and the library sequence) and assigning an identified sequence value based on the sequence annotation and degree of match to each of the sequences in the library;

(b) for some or all entries of the database, tabulating the number of matching identified sequence

values in the library (Although this can be done by human hand from a printout of all entries, we prefer to perform this step using computer software to be described below.), thereby generating a set of final data values or "abundance numbers"; and

(c) if the libraries are different sizes, dividing each abundance number by the total number of sequences in the library, to obtain a relative abundance number for each identified sequence value (i.e., a relative abundance of each gene transcript).

The list of identified sequence values (or genes corresponding thereto) can then be sorted by abundance in the cDNA population. A multitude of additional types of comparisons or dimensions are possible.

For example (to be described below in greater detail), steps (a) and (b) can be repeated for two different libraries (sometimes referred to as a "target" library and a "subtractant" library). Then, for each identified sequence value (or gene transcript), a "ratio" value is obtained by dividing the abundance number (for that identified sequence value) for the target library, by the abundance number (for that identified sequence value) for the subtractant library.

In fact, subtraction may be carried out on multiple libraries. It is possible to add the transcripts from several libraries (for example, three) and then to divide them by another set of transcripts from multiple libraries (again, for example, three). Notation for this operation may be abbreviated as (A+B+C) / (D+E+F), where the capital letters each indicate an entire library. Optionally the abundance numbers of transcripts in the summed libraries may be divided by the total sample size before subtraction.

Unlike standard hybridization technology which permits a single subtraction of two libraries, once one has processed a set or library transcript sequences and stored them in the computer, any number of subtractions can be performed on the library. For example, by this method, ratio values can be obtained by dividing relative abundance

values in a first library by corresponding values in a second library and vice versa.

In variations on step (a), the library consists of nucleotide sequences derived from cDNA clones. Examples of databases which can be searched for areas of homology (similarity) in step (a) include the commercially available databases known as Genbank (NIH) EMBL (European Molecular Biology Labs, Germany), and GENESEQ (Intelligenetics, Mountain View, California).

One homology search algorithm which can be used to implement step (a) is the algorithm described in the paper by D.J. Lipman and W.R. Pearson, entitled "Rapid and Sensitive Protein Similarity Searches," Science, 227:1435 (1985). In this algorithm, the homologous regions are searched in a two-step manner. In the first step, the highest homologous regions are determined by calculating a matching score using a homology score table. The parameter "Ktup" is used in this step to establish the minimum window size to be shifted for comparing two sequences. Ktup also sets the number of bases that must match to extract the highest homologous region among the sequences. In this step, no insertions or deletions are applied and the homology is displayed as an initial (INIT) value.

In the second step, the homologous regions are aligned to obtain the highest matching score by inserting a gap in order to add a probable deleted portion. The matching score obtained in the first step is recalculated using the homology score Table and the insertion score Table to an optimized (OPT) value in the final output.

DNA homologies between two sequences can be examined graphically using the Harr method of constructing dot matrix homology plots (Needleman, S.B. and Wunsch, C.O., J. Mom. Biol 48:443 (1970)). This method produces a two-dimensional plot which can be useful in determining regions of homology versus regions of repetition.

However, in a class of preferred embodiments, step (a) is implemented by processing the library data in the commercially available computer program known as the INHERIT 670 Sequence Analysis System, available from

Applied Biosystems Inc. (Foster City, California), including the software known as the Factura software (also available from Applied Biosystems Inc.). The Factura program preprocesses each library sequence to "edit out" portions thereof which are not likely to be of interest, such as the vector used to prepare the library. Additional sequences which can be edited out or masked (ignored by the search tools) include but are not limited to the polyA tail and repetitive GAG and CCC sequences. A low-end search program can be written to mask out such "low-information" sequences, or programs such as BLAST can ignore the low-information sequences.

In the algorithm implemented by the INHERIT 670 Sequence Analysis System, the Pattern Specification Language (developed by TRW Inc.) is used to determine regions of homology. "There are three parameters that determine how INHERIT analysis runs sequence comparisons: window size, window offset and error tolerance. Window size specifies the length of the segments into which the query sequence is subdivided. Window offset specifies where to start the next segment [to be compared], counting from the beginning of the previous segment. Error tolerance specifies the total number of insertions, deletions and/or substitutions that are tolerated over the specified word length. Error tolerance may be set to any integer between 0 and 6. The default settings are window tolerance=20, window offset=10 and error tolerance=3." INHERIT Analysis Users Manual, pp.2-15. Version 1.0, Applied Biosystems, Inc., October 1991.

Using a combination of these three parameters, a database (such as a DNA database) can be searched for sequences containing regions of homology and the appropriate sequences are scored with an initial value. Subsequently, these homologous regions are examined using dot matrix homology plots to determine regions of homology versus regions of repetition. Smith-Waterman alignments can be used to display the results of the homology search. The INHERIT software can be executed by a Sun computer system programmed with the UNIX operating system.

Search alternatives to INHERIT include the BLAST program, GCG (available from the Genetics Computer Group, WI) and the Dasher program (Temple Smith, Boston University, Boston, MA). Nucleotide sequences can be searched against Genbank, EMBL or custom databases such as GENESEQ (available from Intelligenetics, Mountain View, CA) or other databases for genes. In addition, we have searched some sequences against our own in-house database.

In preferred embodiments, the transcript sequences are analyzed by the INHERIT software for best conformance with a reference gene transcript to assign a sequence identifier and assigned the degree of homology, which together are the identified sequence value and are input into, and further processed by, a Macintosh personal computer (available from Apple) programmed with an "abundance sort and subtraction analysis" computer program (to be described below).

Prior to the abundance sort and subtraction analysis program (also denoted as the "abundance sort" program), identified sequences from the cDNA clones are assigned value (according to the parameters given above) by degree of match according to the following categories: "exact" matches (regions with a high degree of identity), homologous human matches (regions of high similarity, but not "exact" matches), homologous non-human matches (regions of high similarity present in species other than human), or non matches (no significant regions of homology to previously identified nucleotide sequences stored in the form of the database). Alternately, the degree of match can be a numeric value as described below.

With reference again to the step of identifying matches between reference sequences and database entries, protein and peptide sequences can be deduced from the nucleic acid sequences. Using the deduced polypeptide sequence, the match identification can be performed in a manner analogous to that done with cDNA sequences. A protein sequence is used as a query sequence and compared to the previously identified sequences contained in a database such as the Swiss/Prot, PIR and the NBRF Protein database to find homologous proteins. These proteins are

initially scored for homology using a homology score Table (Orcutt, B.C. and Dayoff, M.O. Scoring Matrices, PIR Report MAT - 0285 (February 1985)) resulting in an INIT score. The homologous regions are aligned to obtain the highest matching scores by inserting a gap which adds a probable deleted portion. The matching score is recalculated using the homology score Table and the insertion score Table resulting in an optimized (OPT) score. Even in the absence of knowledge of the proper reading frame of an isolated sequence, the above-described protein homology search may be performed by searching all 3 reading frames.

Peptide and protein sequence homologies can also be ascertained using the INHERIT 670 Sequence Analysis System in an analogous way to that used in DNA sequence homologies. Pattern Specification Language and parameter windows are used to search protein databases for sequences containing regions of homology which are scored with an initial value. Subsequent display in a dot-matrix homology plot shows regions of homology versus regions of repetition. Additional search tools that are available to use on pattern search databases include PLsearch Blocks (available from Henikoff & Henikoff, University of Washington, Seattle), Dasher and GCG. Pattern search databases include, but are not limited to, Protein Blocks (available from Henikoff & Henikoff, University of Washington, Seattle), Brookhaven Protein (available from the Brookhaven National Laboratory, Brookhaven, MA), PROSITE (available from Amos Bairoch, University of Geneva, Switzerland), ProDom (available from Temple Smith, Boston University), and PROTEIN MOTIF FINGERPRINT (available from University of Leeds, United Kingdom).

The ABI Assembler application software, part of the INHERIT DNA analysis system (available from Applied Biosystems, Inc., Foster City, CA), can be employed to create and manage sequence assembly projects by assembling data from selected sequence fragments into a larger sequence. The Assembler software combines two advanced computer technologies which maximize the ability to

assemble sequenced DNA fragments into Assemblages, a special grouping of data where the relationships between sequences are shown by graphic overlap, alignment and statistical views. The process is based on the Meyers-Kececioglu model of fragment assembly (INHERIT™ Assembler User's Manual, Applied Biosystems, Inc., Foster City, CA), and uses graph theory as the foundation of a very rigorous multiple sequence alignment engine for assembling DNA sequence fragments. Other assembly programs that can be used include MEGALIGN (available from DNASTAR Inc., Madison, WI), Dasher and STADEN (available from Roger Staden, Cambridge, England).

Next, with reference to Fig. 2, we describe in more detail the "abundance sort" program which implements above-mentioned "step (b)" to tabulate the number of sequences of the library which match each database entry (the "abundance number" for each database entry).

Fig. 2 is a flow chart of a preferred embodiment of the abundance sort program. A source code listing of this embodiment of the abundance sort program is set forth in Table 5. In the Table 5 implementation, the abundance sort program is written using the FoxBASE programming language commercially available from Microsoft Corporation. Although FoxBASE was the program chosen for the first iteration of this technology, it should not be considered limiting. Many other programming languages, Sybase being a particularly desirable alternative, can also be used, as will be obvious to one with ordinary skill in the art. The subroutine names specified in Fig. 2 correspond to subroutines listed in Table 5.

With reference again to Fig. 2, the "Identified Sequences" are transcript sequences representing each sequence of the library and a corresponding identification of the database entry (if any) which it matches. In other words, the "Identified Sequences" are transcript sequences representing the output of above-discussed "step (a)."

Fig. 3 is a block diagram of a system for implementing the invention. The Fig. 3 system includes library generation unit 2 which generates a library and asserts an

output stream of transcript sequences indicative of the biological sequences comprising the library. Programmed processor 4 receives the data stream output from unit 2 and processes this data in accordance with above-discussed "step (a)" to generate the Identified Sequences. Processor 4 can be a processor programmed with the commercially available computer program known as the INHERIT 670 Sequence Analysis System and the commercially available computer program known as the Factura program (both available from Applied Biosystems Inc.) and with the UNIX operating system.

Still with reference to Fig. 3, the Identified Sequences are loaded into processor 6 which is programmed with the abundance sort program. Processor 6 generates the Final Transcript sequences indicated in both Figs. 2 and 3. Fig. 4 shows a more detailed block diagram of a planned relational computer system, including various searching techniques which can be implemented, along with an assortment of databases to query against.

With reference to Fig. 2, the abundance sort program first performs an operation known as "Tempnum" on the Identified Sequences, to discard all of the Identified Sequences except those which match database entries of selected types. For example, the Tempnum process can select Identified Sequences which represent matches of the following types with database entries (see above for definition): "exact" matches, human "homologous" matches, "other species" matches representing genes present in species other than human), "no" matches (no significant regions of homology with database entries representing previously identified nucleotide sequences), "I" matches (Incyte for not previously known DNA sequences), or "X" matches (matches ESTs in reference database). This eliminates the U, S, M, V, A, R and D sequence (see Table 1 for definitions).

The identified sequence values selected during the "Tempnum" process then undergo a further selection (weeding out) operation known as "Tempred." This operation can, for

example, discard all identified sequence values representing matches with selected database entries.

The identified sequence values selected during the "Tempred" process are then classified according to library, during the "Tempdesig" operation. It is contemplated that the "Identified Sequences" can represent sequences from a single library, or from two or more libraries.

Consider first the case that the identified sequence values represent sequences from a single library. In this case, all the identified sequence values determined during "Tempred" undergo sorting in the "Templib" operation, further sorting in the "Libsort" operation, and finally additional sorting in the "Temptarsort" operation. For example, these three sorting operations can sort the identified sequences in order of decreasing "abundance number" (to generate a list of decreasing abundance numbers, each abundance number corresponding to a unique identified sequence entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list. In this case, the operation identified as "Cruncher" can be bypassed, so that the "Final Data" values are the organized transcript sequences produced during the "Temptarsort" operation.

We next consider the case that the transcript sequences produced during the "Tempred" operation represent sequences from two libraries (which we will denote the "target" library and the "subtractant" library). For example, the target library may consist of cDNA sequences from clones of a diseased cell, while the subtractant library may consist of cDNA sequences from clones of the diseased cell after treatment by exposure to a drug. For another example, the target library may consist of cDNA sequences from clones of a cell type from a young human, while the subtractant library may consist of cDNA sequences from clones of the same cell type from the same human at different ages.

In this case, the "Tempdesig" operation routes all transcript sequences representing the target library for processing in accordance with "Templib" (and then "Libsort" and "Temptarsort"), and routes all transcript sequences representing the subtractant library for processing in accordance with "Tempsub" (and then "Subsort" and "Tempsubsort"). For example, the consecutive "Templib," "Libsort," and "Temptarsort" sorting operations sort identified sequences from the target library in order of decreasing abundance number (to generate a list of decreasing abundance numbers, each abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list. The consecutive "Tempsub," "Subsort," and "Tempsubsort" sorting operations sort identified sequences from the subtractant library in order of decreasing abundance number (to generate a list of decreasing abundance numbers, each abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list.

The transcript sequences output from the "Temptarsort" operation typically represent sorted lists from which a histogram could be generated in which position along one (e.g., horizontal) axis indicates abundance number (of target library sequences), and position along another (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type). Similarly, the transcript sequences output from the "Tempsubsort" operation typically represent sorted lists from which a histogram could be generated in which position along one (e.g., horizontal) axis indicates abundance number (of subtractant library sequences), and position along another (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type).

The transcript sequences (sorted lists) output from the Tempsubsort and Temptarsort sorting operations are combined during the operation identified as "Cruncher." The "Cruncher" process identifies pairs of corresponding target and subtractant abundance numbers (both representing the same identified sequence value), and divides one by the other to generate a "ratio" value for each pair of corresponding abundance numbers, and then sorts the ratio values in order of decreasing ratio value. The data output from the "Cruncher" operation (the Final Transcript sequence in Fig. 2) is typically a sorted list from which a histogram could be generated in which position along one axis indicates the size of a ratio of abundance numbers (for corresponding identified sequence values from target and subtractant libraries) and position along another axis indicates identified sequence value (e.g., gene type).

Preferably, prior to obtaining a ratio between the two library abundance values, the Cruncher operation also divides each ratio value by the total number of sequences in one or both of the target and subtractant libraries. The resulting lists of "relative" ratio values generated by the Cruncher operation are useful for many medical, scientific, and industrial applications. Also preferably, the output of the Cruncher operation is a set of lists, each list representing a sequence of decreasing ratio values for a different selected subset (e.g. protein family) of database entries.

In one example, the abundance sort program of the invention tabulates for a library the numbers of mRNA transcripts corresponding to each gene identified in a database. These numbers are divided by the total number of clones sampled. The results of the division reflect the relative abundance of the mRNA transcripts in the cell type or tissue from which they were obtained. Obtaining this final data set is referred to herein as "gene transcript image analysis." The resulting subtracted data show exactly what proteins and genes are upregulated and downregulated in highly detailed complexity.

### 6.6. HUVEC cDNA LIBRARY

Table 2 is an abundance table listing the various gene transcripts in an induced HUVEC library. The transcripts are listed in order of decreasing abundance. This computerized sorting simplifies analysis of the tissue and speeds identification of significant new proteins which are specific to this cell type. This type of endothelial cell lines tissues of the cardiovascular system, and the more that is known about its composition, particularly in response to activation, the more choices of protein targets become available to affect in treating disorders of this tissue, such as the highly prevalent atherosclerosis.

### 6.7. MONOCYTE-CELL AND MAST-CELL cDNA LIBRARIES

Tables 3 and 4 show truncated comparisons of two libraries. In Tables 3 and 4 the "normal monocytes" are the HMC-1 cells, and the "activated macrophages" are the THP-1 cells pretreated with PMA and activated with LPS. Table 3 lists in descending order of abundance the most abundant gene transcripts for both cell types. With only 15 gene transcripts from each cell type, this table permits quick, qualitative comparison of the most common transcripts. This abundance sort, with its convenient side-by-side display, provides an immediately useful research tool. In this example, this research tool discloses that 1) only one of the top 15 activated macrophage transcripts is found in the top 15 normal monocyte gene transcripts (poly A binding protein); and 2) a new gene transcript (previously unreported in other databases) is relatively highly represented in activated macrophages but is not similarly prominent in normal macrophages. Such a research tool provides researchers with a short-cut to new proteins, such as receptors, cell-surface and intracellular signalling molecules, which can serve as drug targets in commercial drug screening programs. Such a tool could save considerable time over that consumed by a hit and miss discovery program aimed at identifying important proteins in and around cells, because those proteins carrying out everyday cellular functions and

represented as steady state mRNA are quickly eliminated from further characterization.

This illustrates how the gene transcript profiles change with altered cellular function. Those skilled in the art know that the biochemical composition of cells also changes with other functional changes such as cancer, including cancer's various stages, and exposure to toxicity. A gene transcript subtraction profile such as in Table 3 is useful as a first screening tool for such gene expression and protein studies.

### 6.8. SUBTRACTION ANALYSIS OF NORMAL MONOCYTE-CELL AND ACTIVATED MONOCYTE CELL cDNA LIBRARIES

Once the cDNA data are in the computer, the computer program as disclosed in Table 5 was used to obtain ratios of all the gene transcripts in the two libraries discussed in Example 6.7, and the gene transcripts were sorted by the descending values of their ratios. If a gene transcript is not represented in one library, that gene transcript's abundance is unknown but appears to be less than 1. As an approximation -- and to obtain a ratio, which would not be possible if the unrepresented gene were given an abundance of zero -- genes which are represented in only one of the two libraries are assigned an abundance of 1/2. Using 1/2 for unrepresented clones increases the relative importance of "turned-on" and "turned-off" genes, whose products would be drug candidates. The resulting print-out is called a subtraction table and is an extremely valuable screening method, as is shown by the following data.

Table 4 is a subtraction table, in which the normal monocyte library was electronically "subtracted" from the activated macrophage library. This table highlights most effectively the changes in abundance of the gene transcripts by activation of macrophages. Even among the first 20 gene transcripts listed, there are several unknown gene transcripts. Thus, electronic subtraction is a useful tool with which to assist researchers in identifying much more quickly the basic biochemical changes between two cell types. Such a tool can save universities and pharmaceutical companies which spend billions of dollars on

research valuable time and laboratory resources at the early discovery stage and can speed up the drug development cycle, which in turn permits researchers to set up drug screening programs much earlier. Thus, this research tool provides a way to get new drugs to the public faster and more economically.

Also, such a subtraction table can be obtained for patient diagnosis. An individual patient sample (such as monocytes obtained from a biopsy or blood sample) can be compared with data provided herein to diagnose conditions associated with macrophage activation.

Table 4 uncovered many new gene transcripts (labeled Incyte clones). Note that many genes are turned on in the activated macrophage (i.e., the monocyte had a 0 in the bgfreq column). This screening method is superior to other screening techniques, such as the western blot, which are incapable of uncovering such a multitude of discrete new gene transcripts.

The subtraction-screening technique has also uncovered a high number of cancer gene transcripts (oncogenes rho, ETS2, rab-2 ras, YPT1-related, and acute myeloid leukemia mRNA) in the activated macrophage. These transcripts may be attributed to the use of immortalized cell lines and are inherently interesting for that reason. This screening technique offers a detailed picture of upregulated transcripts including oncogenes, which helps explain why anti-cancer drugs interfere with the patient's immunity mediated by activated macrophages. Armed with knowledge gained from this screening method, those skilled in the art can set up more targeted, more effective drug screening programs to identify drugs which are differentially effective against 1) both relevant cancers and activated macrophage conditions with the same gene transcript profile; 2) cancer alone; and 3) activated macrophage conditions.

Smooth muscle senescent protein (22 kd) was upregulated in the activated macrophage, which indicates that it is a candidate to block in controlling inflammation.

### 6.9. SUBTRACTION ANALYSIS OF NORMAL LIVER CELLS AND HEPATITIS INFECTED LIVER CELL cDNA LIBRARIES

In this example, rats are exposed to hepatitis virus and maintained in the colony until they show definite signs of hepatitis. Of the rats diagnosed with hepatitis, one half of the rats are treated with a new anti-hepatitis agent (AHA). Liver samples are obtained from all rats before exposure to the hepatitis virus and at the end of AHA treatment or no treatment. In addition, liver samples can be obtained from rats with hepatitis just prior to AHA treatment.

The liver tissue is treated as described in Examples 6.2 and 6.3 to obtain mRNA and subsequently to sequence cDNA. The cDNA from each sample are processed and analyzed for abundance according to the computer program in Table 5. The resulting gene transcript images of the cDNA provide detailed pictures of the baseline (control) for each animal and of the infected and/or treated state of the animals. cDNA data for a group of samples can be combined into a group summary gene transcript profile for all control samples, all samples from infected rats and all samples from AHA-treated rats.

Subtractions are performed between appropriate individual libraries and the grouped libraries. For individual animals, control and post-study samples can be subtracted. Also, if samples are obtained before and after AHA treatment, that data from individual animals and treatment groups can be subtracted. In addition, the data for all control samples can be pooled and averaged. The control average can be subtracted from averages of both post-study AHA and post-study non-AHA cDNA samples. If pre- and post-treatment samples are available, pre- and post-treatment samples can be compared individually (or electronically averaged) and subtracted.

These subtraction tables are used in two general ways. First, the differences are analyzed for gene transcripts which are associated with continuing hepatic deterioration or healing. The subtraction tables are tools to isolate the effects of the drug treatment from the underlying basic pathology of hepatitis. Because hepatitis affects many

parameters, additional liver toxicity has been difficult to detect with only blood tests for the usual enzymes. The gene transcript profile and subtraction provides a much more complex biochemical picture which researchers have needed to analyze such difficult problems.

Second, the subtraction tables provide a tool for identifying clinical markers, individual proteins or other biochemical determinants which are used to predict and/or evaluate a clinical endpoint, such as disease, improvement due to the drug, and even additional pathology due to the drug. The subtraction tables specifically highlight genes which are turned on or off. Thus, the subtraction tables provide a first screen for a set of gene transcript candidates for use as clinical markers. Subsequently, electronic subtractions of additional cell and tissue libraries reveal which of the potential markers are in fact found in different cell and tissue libraries. Candidate gene transcripts found in additional libraries are removed from the set of potential clinical markers. Then, tests of blood or other relevant samples which are known to lack and have the relevant condition are compared to validate the selection of the clinical marker. In this method, the particular physiologic function of the protein transcript need not be determined to qualify the gene transcript as a clinical marker.

## 6.10. ELECTRONIC NORTHERN BLOT

One limitation of electronic subtraction is that it is difficult to compare more than a pair of images at once. Once particular individual gene products are identified as relevant to further study (via electronic subtraction or other methods), it is useful to study the expression of single genes in a multitude of different tissues. In the lab, the technique of "Northern" blot hybridization is used for this purpose. In this technique, a single cDNA, or a probe corresponding thereto, is labeled and then hybridized against a blot containing RNA samples prepared from a multitude of tissues or cell types. Upon autoradiography,

the pattern of expression of that particular gene, one at a time, can be quantitated in all the included samples.

In contrast, a further embodiment of this invention is the computerized form of this process, termed here "electronic northern blot." In this variation, a single gene is queried for expression against a multitude of prepared and sequenced libraries present within the database. In this way, the pattern of expression of any single candidate gene can be examined instantaneously and effortlessly. More candidate genes can thus be scanned, leading to more frequent and fruitfully relevant discoveries. The computer program included as Table 5 includes a program for performing this function, and Table 6 is a partial listing of entries of the database used in the electronic northern blot analysis.

### 6.11. PHASE I CLINICAL TRIALS

Based on the establishment of safety and effectiveness in the above animal tests, Phase I clinical tests are undertaken. Normal patients are subjected to the usual preliminary clinical laboratory tests. In addition, appropriate specimens are taken and subjected to gene transcript analysis. Additional patient specimens are taken at predetermined intervals during the test. The specimens are subjected to gene transcript analysis as described above. In addition, the gene transcript changes noted in the earlier rat toxicity study are carefully evaluated as clinical markers in the followed patients. Changes in the gene transcript analyses are evaluated as indicators of toxicity by correlation with clinical signs and symptoms and other laboratory results. In addition, subtraction is performed on individual patient specimens and on averaged patient specimens. The subtraction analysis highlights any toxicological changes in the treated patients. This is a highly refined determinant of toxicity. The subtraction method also annotates clinical markers. Further subgroups can be analyzed by subtraction analysis, including, for example, 1) segregation by

occurrence and type of adverse effect; and 2) segregation by dosage.

## 6.12. GENE TRANSCRIPT IMAGING ANALYSIS IN CLINICAL STUDIES

A gene transcript imaging analysis (or multiple gene transcript imaging analyses) is a useful tool in other clinical studies. For example, the differences in gene transcript imaging analyses before and after treatment can be assessed for patients on placebo and drug treatment. This method also effectively screens for clinical markers to follow in clinical use of the drug.

## 6.13. COMPARATIVE GENE TRANSCRIPT ANALYSIS BETWEEN SPECIES

The subtraction method can be used to screen cDNA libraries from diverse sources. For example, the same cell types from different species can be compared by gene transcript analysis to screen for specific differences, such as in detoxification enzyme systems. Such testing aids in the selection and validation of an animal model for the commercial purpose of drug screening or toxicological testing of drugs intended for human or animal use. When the comparison between animals of different species is shown in columns for each species, we refer to this as an interspecies comparison, or zoo blot.

Embodiments of this invention may employ databases such as those written using the FoxBASE programming language commercially available from Microsoft Corporation. Other embodiments of the invention employ other databases, such as a random peptide database, a polymer database, a synthetic oligomer database, or a oligonucleotide database of the type described in U.S. Patent 5,270,170, issued December 14, 1993 to Cull, et al., PCT International Application Publication No. WO 9322684, published November 11, 1993, PCT International Application Publication No. WO 9306121, published April 1, 1993, or PCT International Application Publication No. WO 9119818, published December 26, 1991. These four references (whose text is incorporated herein by reference) include teaching which

may be applied in implementing such other embodiments of the present invention.

All references referred to in the preceding text are hereby expressly incorporated by reference herein.

Various modifications and variations of the described method and system of the invention will be apparent to those skilled in the art without departing from the scope and spirit of the invention. Although the invention has been described in connection with specific preferred embodiments, it should be understood that the invention as claimed should not be unduly limited to such specific embodiments.

TABLE 1

**Designations (D)**

E = Exact
H = Homologous
O = Other species
N = No match
D = Noncoding gene
U = Nonreadable
R = Repetitive DNA
A = Poly-A only
V = Vector only
M = Mitochondrial DNA
S = Skip
I = Match Incyte clone
X = EST match

**Library (L)**

U = U937
M = HMC
T = THP-1
H = HUVEC
S = Spleen
L = Lung
Y = T & B cell
A = Adenoid

**Distribution (F)**

C = Non-specific
P = Cell/tissue specific
U = Unknown

**Species (S)**

H = Human
A = Ape
P = Pig
D = Dog
V = Bovine
B = Rabbit
R = Rat
M = Mouse
S = Hamster
C = Chicken
F = Amphibian
I = Invertebrate
Z = Protozoan
G = Fungi

**Localization (Z)**

N = Nuclear
C = Cytoplasmic
K = Cytoskeleton
E = Cell surface
Z = Intracellular memb
M = Mitochondrial
S = Secreted
U = Unknown
X = Other

**Status (I)**

0 = No current interest
1 = Do primary analysis
2 = Primary analysis done
3 = Full length sequence
4 = Secondary analysis
5 = Tissue northern
6 = Obtain full length

**Function (R)**

T = Translation
L = Protein processing
R = Ribosomal protein
O = *Oncogene*
  G = GTP binding ptn
  V = Viral element
  Y = Kinase/phosphatase
  A = Tumor antigen related
I = *Binding proteins*
  D = NA-binding /transcription
  B = Surface molecule/receptor
  C = $Ca^{++}$ binding protein
  S = Ligands/effectors
H = Stress response protein
E = *Enzyme*
  F = Ferroprotein
  P = Protease/inhibitor
  Z = Oxidative phosphorylation
  Q = Sugar metabolism
  M = Amino acid metabolism
  N = Nucleic acid metabolism
  W = Lipid metabolism
  K = Structural
  X = Other
  U = unknown

TABLE 2

Clone numbers 15000 through 20000
Libraries: HUVEC
Arranged by ABUNDANCE
Total clones analyzed: 5000

319 genes, for a total of 1713 Clones

| | number | N | c | entry | s | descriptor |
|---|---|---|---|---|---|---|
| 1 | 15365 | 67 | | HSRPL41 | | Riboptn L41 |
| 2 | 15004 | 65 | | NCY015004 | | INCYTE 015004 |
| 3 | 15638 | 63 | | NCY015638 | | INCYTE 015638 |
| 4 | 15390 | 50 | | NCY015390 | | INCYTE 015390 |
| 5 | 15193 | 47 | | HSFIB1 | | Fibronectin |
| 6 | 15220 | 47 | | RRRPL9 | R | Riboptn L9 |
| 7 | 15280 | 47 | | NCY015280 | | INCYTE 015280 |
| 8 | 15583 | 33 | | M62060 | | EST HHCH09 (IGR) |
| 9 | 15662 | 31 | | HSACTCGR | | Actin, gamma |
| 10 | 15026 | 29 | | NCY015026 | | INCYTE 015026 |
| 11 | 15279 | 24 | | HSEF1AR | | Elf 1-alpha |
| 12 | 15027 | 23 | | NCY015027 | | INCYTE 015027 |
| 13 | 15033 | 20 | | NCY015033 | | INCYTE 015033 |
| 14 | 15198 | 20 | | NCY015198 | | INCYTE 015198 |
| 15 | 15809 | 20 | | HSCOLL1 | | Collagenase |
| 16 | 15221 | 19 | | NCY015221 | | INCYTE 015221 |
| 17 | 15263 | 19 | | NCY015263 | | INCYTE 015263 |
| 18 | 15290 | 19 | | NCY015290 | | INCYTE 015290 |
| 19 | 15350 | 18 | | NCY015350 | | INCYTE 015350 |
| 20 | 15030 | 17 | | NCY015030 | | INCYTE 015030 |
| 21 | 15234 | 17 | | NCY015234 | | INCYTE 015234 |
| 22 | 15459 | 16 | | NCY015459 | | INCYTE 015459 |
| 23 | 15353 | 15 | | NCY015353 | | INCYTE 015353 |
| 24 | 15378 | 15 | | S76965 | | Ptn kinase inhib |
| 25 | 15255 | 14 | | HUMTHYB4 | | Thymosin beta-4 |
| 26 | 15401 | 14 | | HSLIPCR | | Lipocortin I |
| 27 | 15425 | 14 | | HSPOLYAB | | Poly-A bp |
| 28 | 18212 | 14 | | HUMTHYMA | | Thymosin, alpha |
| 29 | 18216 | 14 | | HSMRP1 | | Motility relat ptn; MRP-1;CD-9 |
| 30 | 15189 | 13 | | HS18D | | Interferon induc ptn 1-8D |
| 31 | 15031 | 12 | | HUMFKBP | | FK506 bp |
| 32 | 15306 | 12 | | HSH2AZ | | Histone H2A |
| 33 | 15621 | 12 | | HUMLEC | | Lectin, B-galbp, 14kDa |
| 34 | 15789 | 11 | | NCY015789 | | INCYTE 015789 |
| 35 | 16578 | 11 | | HSRPS11 | | Riboptn S11 |
| 36 | 16632 | 11 | | M61984 | | EST HHCA13 (IGR) |
| 37 | 18314 | 11 | | NCY018314 | | INCYTE 018314 |
| 38 | 15367 | 10 | | NCY015367 | | INCYTE 015367 |
| 39 | 15415 | 10 | | HSIFNIN1 | | interferon induc mRNA |
| 40 | 15633 | 10 | | HSLDHAR | | Lactate dehydrogenase |
| 41 | 15813 | 10 | | CHKNMHCB | | C Myosin heavy chain B |
| 42 | 18210 | 10 | | NCY018210 | | INCYTE 018210 |
| 43 | 18233 | 10 | | HSRPII140 | | RNA polymerase II |
| 44 | 18996 | 10 | | NCY018996 | | INCYTE 018996 |
| 45 | 15088 | 9 | | HUMFERL | | Ferritin, light chain |
| 46 | 15714 | 9 | | NCY015714 | | INCYTE 015714 |
| 47 | 15720 | 9 | | NCY015720 | | INCYTE 015720 |
| 48 | 15863 | 9 | | NCY015863 | | INCYTE 015863 |
| 49 | 16121 | 9 | | HSET | | Endothelin |
| 50 | 18252 | 9 | | NCY018252 | | INCYTE 018252 |
| 51 | 15351 | 8 | | HUMALBP | | Lipid bp, adipocyte |
| 52 | 15370 | 8 | | NCY015370 | | INCYTE 015370 |

TABLE 2 Con't

| | number | N | c | entry | s | descriptor |
|---|---|---|---|---|---|---|
| 53 | 15670 | 8 | | BTCIASHI | V | NADH-ubiq oxidoreductase |
| 54 | 15795 | 8 | | NCY015795 | | INCYTE 015795 |
| 55 | 16245 | 8 | | NCY016245 | | INCYTE 016245 |
| 56 | 18262 | 8 | | NCY018262 | | INCYTE 018262 |
| 57 | 18321 | 8 | | HSRPL17 | | Riboptn L17 |
| 58 | 15126 | 7 | | XLRPL1BRF | | Riboptn L1 |
| 59 | 15133 | 7 | | HSAC07 | | Actin, beta |
| 60 | 15245 | 7 | | NCY015245 | | INCYTE 015245 |
| 61 | 15288 | 7 | | NCY015288 | | INCYTE 015288 |
| 62 | 15294 | 7 | | HSGAPDR | | G-3-PD |
| 63 | 15442 | 7 | | HUMLAMB | | Laminin receptor, 54kDa |
| 64 | 15485 | 7 | | HSNGMRNA | | Uracil DNA glycosylase |
| 65 | 16646 | 7 | | NCY016646 | | INCYTE 016646 |
| 66 | 18003 | 7 | | HUMPAIA | | Plsmnogen activ gene |
| 67 | 15032 | 6 | | HUMUB | | Ubiquitin |
| 68 | 15267 | 6 | | HSRPS8 | | Riboptn S8 |
| 69 | 15295 | 6 | | NCY015295 | | INCYTE 015295 |
| 70 | 15458 | 6 | | RNRPS10R | R | Riboptn S10 |
| 71 | 15832 | 6 | | RSGALEM | R | UDP-galactose epimerase |
| 72 | 15928 | 6 | | HUMAPOJ | | Apolipoptn J |
| 73 | 16598 | 6 | | HUMTBBM40 | | Tubulin, beta |
| 74 | 18218 | 6 | | NCY018218 | | INCYTE 018218 |
| 75 | 18499 | 6 | | HSP27 | | Hydrophobic ptn p27 |
| 76 | 18963 | 6 | | NCY018963 | | INCYTE 018963 |
| 77 | 18997 | 6 | | NCY018997 | | INCYTE 018997 |
| 78 | 15432 | 5 | | HSAGALAR | | Galactosidase A, alpha |
| 79 | 15475 | 5 | | NCY015475 | | INCYTE 015475 |
| 80 | 15721 | 5 | | NCY015721 | | INCYTE 015721 |
| 81 | 15865 | 5 | | NCY015865 | | INCYTE 015865 |
| 82 | 16270 | 5 | | NCY016270 | | INCYTE 016270 |
| 83 | 16886 | 5 | | NCY016886 | | INCYTE 016886 |
| 84 | 18500 | 5 | | NCY018500 | | INCYTE 018500 |
| 85 | 18503 | 5 | | NCY018503 | | INCYTE 018503 |
| 86 | 19672 | 5 | | RRRPL34 | R | Riboptn L34 |
| 87 | 15086 | 4 | | XLRPL1AR | F | Riboptn L1a |
| 88 | 15113 | 4 | | HUMIFNWRS | | tRNA synthetase, trp |
| 89 | 15242 | 4 | | NCY015242 | | INCYTE 015242 |
| 90 | 15249 | 4 | | NCY015249 | | INCYTE 015249 |
| 91 | 15377 | 4 | | NCY015377 | | INCYTE 015377 |
| 92 | 15407 | 4 | | NCY015407 | | INCYTE 015407 |
| 93 | 15473 | 4 | | NCY015473 | | INCYTE 015473 |
| 94 | 15588 | 4 | | HSRPS12 | | Riboptn S12 |
| 95 | 15684 | 4 | | HSEF1G | | Elf 1-gamma |
| 96 | 15782 | 4 | | NCY015782 | | INCYTE 015782 |
| 97 | 15916 | 4 | | HSRPS18 | | Riboptn S18 |
| 98 | 15930 | 4 | | NCY015930 | | INCYTE 015930 |
| 99 | 16108 | 4 | | NCY016108 | | INCYTE 016108 |
| 100 | 16133 | 4 | | NCY016133 | | INCYTE 016133 |

# NORMAL MONONCYTE VS. ACTIVATED MACROPHAGE

## *Top 15 Most Abundant Genes*

| | NORMAL | ACTIVATED |
|---|---|---|
| 1 | Elongation factor-1 alpha | Interleukin-1 beta |
| 2 | Ribosomal phosphoprotein | Macrophage inflammatory protein-1 |
| 3 | Ribosomal protein S8 homolog | Interleukin-8 |
| 4 | Beta-Globin | Lymphocyte activation gene |
| 5 | Ferritin H chain | Elongation factor-1 alpha |
| 6 | Ribosomal protein L7 | Beta actin |
| 7 | Nucleoplasmin | Rantes T-cell specific protein |
| 8 | Ribosomal protein S20 homolog | Poly A binding protein |
| 9 | Transferrin receptor | Osteopontin; nephropontin |
| 10 | Poly-A binding protein | Tumor Necrosis Factor-alpha |
| 11 | Translationally controlled tumor ptn | INCYTE clone 011050 |
| 12 | Ribosomal protein S25 | Cu/Zn superoxide dismutase |
| 13 | Signal recognition particle SRP9 | Adenylate cyclase (yeast homolog) |
| 14 | Histone H2A.Z | NGF-related B cell activation molecule |
| 15 | Ribosomal protein Ke-3 | Protease Nexin-1, glial-derived |

TABLE 3

TABLE 4

Libraries: THP-1
Subtracting: HMC
Sorted by ABUNDANCE
Total clones analyzed: 7375

1057 genes, for a total of 2151 clones

| number | entry | s | descriptor | bgfreq | rfend | ratio |
|---|---|---|---|---|---|---|
| 10022 | HUMIL1 | | IL 1-beta | 0 | 131 | 262.00 |
| 10036 | HSMDNCF | | IL-8 | 0 | 119 | 238.00 |
| 10089 | HSLAG1CDN | | Lymphocyte activ gene | 0 | 71 | 142.00 |
| 10060 | HUMTCSM | | RANTES | 0 | 23 | 46.000 |
| 10003 | HUMMIP1A | | MIP-1 | 3 | 121 | 40.333 |
| 10689 | HSOP | | Osteopontin | 0 | 20 | 40.000 |
| 11050 | NCY011050 | | INCYTE 011050 | 0 | 17 | 34.000 |
| 10937 | HSTNFR | | TNF-alpha | 0 | 17 | 34.000 |
| 10176 | HSSOD | | Superoxide dismutase | 0 | 14 | 28.000 |
| 10886 | HSCDW40 | | B-cell activ,NGF-relat | 0 | 10 | 20.000 |
| 10186 | HUMAPR | | Early resp PMA-induc | 0 | 9 | 18.000 |
| 10967 | HUMGDN | | PN-1, glial-deriv | 0 | 9 | 18.000 |
| 11353 | NCY011353 | | INCYTE 011353 | 0 | 8 | 16.000 |
| 10298 | NCY010298 | | INCYTE 010298 | 0 | 7 | 14.000 |
| 10215 | HUM4COLA | | Collagenase, type IV | 0 | 6 | 12.000 |
| 10276 | NCY010276 | | INCYTE 010276 | 0 | 6 | 12.000 |
| 10488 | NCY010488 | | INCYTE 010488 | 0 | 6 | 12.000 |
| 11138 | NCY011138 | | INCYTE 011138 | 0 | 6 | 12.000 |
| 10037 | HUMCAPPRO | | Adenylate cyclase | 1 | 10 | 10.000 |
| 10840 | HUMADCY | | Adenylate cyclase | 0 | 5 | 10.000 |
| 10672 | HSCD44E | | Cell adhesion glptn | 0 | 5 | 10.000 |
| 12837 | HUMCYCLOX | | Cyclooxygenase-2 | 0 | 5 | 10.000 |
| 10001 | NCY010001 | | INCYTE 010001 | 0 | 5 | 10.000 |
| 10005 | NCY010005 | | INCYTE 010005 | 0 | 5 | 10.000 |
| 10294 | NCY010294 | | INCYTE 010294 | 0 | 5 | 10.000 |
| 10297 | NCY010297 | | INCYTE 010297 | 0 | 5 | 10.000 |
| 10403 | NCY010403 | | INCYTE 010403 | 0 | 5 | 10.000 |
| 10699 | NCY010699 | | INCYTE 010699 | 0 | 5 | 10.000 |
| 10966 | NCY010966 | | INCYTE 010966 | 0 | 5 | 10.000 |
| 12092 | NCY012092 | | INCYTE 012092 | 0 | 5 | 10.000 |
| 12549 | HSRHOB | | Oncogene rho | 0 | 5 | 10.000 |
| 10691 | HUMARF1BA | | ADP-ribosylation fctr | 0 | 4 | 8.000 |
| 12106 | HSADSS | | Adenylosuccinate synthetase | 0 | 4 | 8.000 |
| 10194 | HSCATHL | | Cathepsin L | 0 | 4 | 8.000 |
| 10479 | CLMCYCA | I | Cyclin A | 0 | 4 | 8.000 |
| 10031 | NCY010031 | | INCYTE 010031 | 0 | 4 | 8.000 |
| 10203 | NCY010203 | | INCYTE 010203 | 0 | 4 | 8.000 |
| 10288 | NCY010288 | | INCYTE 010288 | 0 | 4 | 8.000 |
| 10372 | NCY010372 | | INCYTE 010372 | 0 | 4 | 8.000 |
| 10471 | NCY010471 | | INCYTE 010471 | 0 | 4 | 8.000 |
| 10484 | NCY010484 | | INCYTE 010484 | 0 | 4 | 8.000 |
| 10859 | NCY010859 | | INCYTE 010859 | 0 | 4 | 8.000 |
| 10890 | NCY010890 | | INCYTE 010890 | 0 | 4 | 8.000 |
| 11511 | NCY011511 | | INCYTE 011511 | 0 | 4 | 8.000 |
| 11868 | NCY011868 | | INCYTE 011868 | 0 | 4 | 8.000 |
| 12820 | NCY012820 | | INCYTE 012820 | 0 | 4 | 8.000 |
| 10133 | HSI1RAP | | IL-1 antagonist | 0 | 4 | 8.000 |
| 10516 | HUMP2A | | Phosphatase, regul 2A | 0 | 4 | 8.000 |
| 11063 | HUMB94 | | TNF-induc response | 0 | 4 | 8.000 |
| 11140 | HSHB15RNA | | HB15 gene; new Ig | 0 | 3 | 6.000 |
| 10788 | NCY001713 | | INCYTE 001713 | 0 | 3 | 6.000 |
| 10033 | NCY010033 | | INCYTE 010033 | 0 | 3 | 6.000 |
| 10035 | NCY010035 | | INCYTE 010035 | 0 | 3 | 6.000 |
| 10084 | NCY010084 | | INCYTE 010084 | 0 | 3 | 6.000 |
| 10236 | NCY010236 | | INCYTE 010236 | 0 | 3 | 6.000 |
| 10383 | NCY010383 | | INCYTE 010383 | 0 | 3 | 6.000 |

TABLE 4 Con't

| number | entry | s descriptor | bgfreq | rfend | ratio |
|---|---|---|---|---|---|
| 10450 | NCY010450 | INCYTE 010450 | 0 | 3 | 6.000 |
| 10470 | NCY010470 | INCYTE 010470 | 0 | 3 | 6.000 |
| 10504 | NCY010504 | INCYTE 010504 | 0 | 3 | 6.000 |
| 10507 | NCY010507 | INCYTE 010507 | 0 | 3 | 6.000 |
| 10598 | NCY010598 | INCYTE 010598 | 0 | 3 | 6.000 |
| 10779 | NCY010779 | INCYTE 010779 | 0 | 3 | 6.000 |
| 10909 | NCY010909 | INCYTE 010909 | 0 | 3 | 6.000 |
| 10976 | NCY010976 | INCYTE 010976 | 0 | 3 | 6.000 |
| 10985 | NCY010985 | INCYTE 010985 | 0 | 3 | 6.000 |
| 11052 | NCY011052 | INCYTE 011052 | 0 | 3 | 6.000 |
| 11068 | NCY011068 | INCYTE 011068 | 0 | 3 | 6.000 |
| 11134 | NCY011134 | INCYTE 011134 | 0 | 3 | 6.000 |
| 11136 | NCY011136 | INCYTE 011136 | 0 | 3 | 6.000 |
| 11191 | NCY011191 | INCYTE 011191 | 0 | 3 | 6.000 |
| 11219 | NCY011219 | INCYTE 011219 | 0 | 3 | 6.000 |
| 11386 | NCY011386 | INCYTE 011386 | 0 | 3 | 6.000 |
| 11403 | NCY011403 | INCYTE 011403 | 0 | 3 | 6.000 |
| 11460 | NCY011460 | INCYTE 011460 | 0 | 3 | 6.000 |
| 11618 | NCY011618 | INCYTE 011618 | 0 | 3 | 6.000 |
| 11686 | NCY011686 | INCYTE 011686 | 0 | 3 | 6.000 |
| 12021 | NCY012021 | INCYTE 012021 | 0 | 3 | 6.000 |
| 12025 | NCY012025 | INCYTE 012025 | 0 | 3 | 6.000 |
| 12320 | NCY012320 | INCYTE 012320 | 0 | 3 | 6.000 |
| 12330 | NCY012330 | INCYTE 012330 | 0 | 3 | 6.000 |
| 12853 | NCY012853 | INCYTE 012853 | 0 | 3 | 6.000 |
| 14386 | NCY014386 | INCYTE 014386 | 0 | 3 | 6.000 |
| 14391 | NCY014391 | INCYTE 014391 | 0 | 3 | 6.000 |

TABLE 5

```
* Master menu for SUBTRACTION output
SET TALK OFF
SET SAFETY OFF
SET EXACT ON
SET TYPEAHEAD TO 0
CLEAR
SET DEVICE TO SCREEN
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE '          ' TO Target1
STORE '          ' TO Target2
STORE '          ' TO Target3
STORE '          ' TO Object1
STORE '          ' TO Object2
STORE '          ' TO Object3
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO PTF
STORE 1 TO BAIL
DO WHILE .T.
* Program.: Subtraction 2.fmt
* Date....: 10/11/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes...: Format file Subtraction 2
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",9 COLOR 0,0,0,
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,24610,-1,8947
@ PIXELS 27,134 SAY "Subtraction Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Exact " SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Homologous" SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Other spc" SIZE 15,84
@ PIXELS 90,152 SAY "Matches:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "@*C Incyte" SIZE 15,65 CO
@ PIXELS 252,137 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,236 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,35 SAY "Include clones:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,215 SAY "->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",12 PICTURE "@*C Print to file" SIZE 15,9
@ PIXELS 90,9 TO 181,109 STYLE 3871 COLOR 0,0,-1,-25500,-1,-1
@ PIXELS 90,288 TO 181,397 STYLE 3871 COLOR 0,0,-1,-25500,-1,-1
@ PIXELS 81,296 SAY "Background:" STYLE 65536 FONT "Geneva",270 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 45,135 GET ANAL STYLE 65536 FONT "Chicago",12 PICTURE "@*R Overall;Function" SIZE 4
@ PIXELS 81,26 SAY "Target:" STYLE 65536 FONT "Geneva",270 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 108,20 GET target1 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 135,20 GET target2 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 162,20 GET target3 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 108,299 GET object1 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 135,299 GET object2 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 162,299 GET object3 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 276,324 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Run;Bail out" SIZE 4112
*
* EOF: Subtraction 2.fmt
READ
  IF Bail=2
  CLEAR
  CLOSE DATABASES
  USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
  SET SAFETY ON
  SCREEN 1 OFF
  RETURN
```

```
   ENDIF
  STORE VAL(SYS(2)) TO STARTIME
  STORE UPPER(Target1) TO Target1
  STORE UPPER(Target2) TO Target2
  STORE UPPER(Target3) TO Target3
  STORE UPPER(Object1) TO Object1
  STORE UPPER(Object2) TO Object2
  STORE UPPER(Object3) TO Object3
  clear
  SET TALK ON
  GAP = TERMINATE-INITIATE+1
  GO INITIATE
  COPY NEXT GAP FIELDS NUMBER,library,D,F,Z,R,ENTRY,S,DESCRIPTOR,START,RFEND,I TO TEMPNUM
  USE TEMPNUM
  COUNT TO TOT
  COPY TO TEMPRED FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='I'
  USE TEMPRED

  IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
  COPY TO TEMPDESIG
  ELSE
  COPY STRUCTURE TO TEMPDESIG
  USE TEMPDESIG
    IF Ematch=1
    APPEND FROM TEMPNUM FOR D='E'
    ENDIF
    IF Hmatch=1
    APPEND FROM TEMPNUM FOR D='H'
    ENDIF
    IF Omatch=1
    APPEND FROM TEMPNUM FOR D='O'
    ENDIF
    IF Imatch=1
    APPEND FROM TEMPNUM FOR D='I'.OR.D='X'
*.OR.D='N'
    ENDIF
ENDIF
COUNT TO STARTOT

COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
  APPEND FROM TEMPDESIG FOR library=UPPER(target1)
  IF target2<>'
  APPEND FROM TEMPDESIG FOR library=UPPER(target2)
  ENDIF
  IF target3<>'
  APPEND FROM TEMPDESIG FOR library=UPPER(target3)
  ENDIF
COUNT TO ANALTOT

USE TEMPDESIG
COPY STRUCTURE TO TEMPSUB
USE TEMPSUB
  APPEND FROM TEMPDESIG FOR library=UPPER(Object1)
  IF target2<>'
  APPEND FROM TEMPDESIG FOR library=UPPER(Object2)
  ENDIF
  IF target3<>'
  APPEND FROM TEMPDESIG FOR library=UPPER(Object3)
  ENDIF
COUNT TO SUBTRACTOT
SET TALK OFF
*************************************************************************************************
* COMPRESSION SUBROUTINE A
? 'COMPRESSING QUERY LIBRARY'
USE TEMPLIB
```

```
SORT ON ENTRY,NUMBER TO LIBSORT
USE LIBSORT
COUNT TO IDGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= IDGENE
  PACK
  COUNT TO AUNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
STORE D TO DESIGB
  IF TESTA = TESTB.AND.DESIGA=DESIGB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
SORT ON RFEND/D,NUMBER TO TEMPTARSORT.
USE TEMPTARSORT
*REPLACE ALL START WITH RFEND/IDGENE*10000
COUNT TO TEMPTARCO
**********
* COMPRESSION SUBROUTINE B
? 'COMPRESSING TARGET LIBRARY'
USE TEMPSUB
SORT ON ENTRY,NUMBER TO SUBSORT
USE SUBSORT
COUNT TO SUBGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= SUBGENE
  PACK
  COUNT TO SUNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
STORE D TO DESIGB
  IF TESTA = TESTB.AND.DESIGA=DESIGB
```

```
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
 GO MARK1
 REPLACE RFEND WITH DUP
 MARK1 = MARK1+DUP
 SW=1
 LOOP
 ENDDO TEST
 LOOP
 ENDDO ROLL
 SORT ON RFEND/D,NUMBER TO TEMPSUBSORT
 USE TEMPSUBSORT
 *REPLACE ALL START WITH RFEND/IDGENE*10000
 COUNT TO TEMPSUBCO
 ************************************************************************************************
 *FUSION ROUTINE
 ? 'SUBTRACTING LIBRARIES' -
 USE SUBTRACTION
 COPY STRUCTURE TO CRUNCHER
 SELECT 2
 USE TEMPSUBSORT
 SELECT 1
 USE CRUNCHER
 APPEND FROM TEMPTARSORT
 COUNT TO BAILOUT
 MARK = 0

 DO WHILE .T.
 SELECT 1
 MARK = MARK+1
   IF MARK>BAILOUT
   EXIT
   ENDIF
 GO MARK
 STORE ENTRY TO SCANNER
 SELECT 2
 LOCATE FOR ENTRY=SCANNER
 IF FOUND()
 STORE RFEND TO BIT1
 STORE RFEND TO BIT2
 ELSE
 STORE 1/2 TO BIT1
 STORE 0 TO BIT2
 ENDIF
 SELECT 1
 REPLACE BGFREQ WITH BIT2
 REPLACE ACTUAL WITH BIT1
 LOOP
 ENDDO

 SELECT 1
 REPLACE ALL RATIO WITH RFEND/ACTUAL
 ? 'DOING FINAL SORT BY RATIO'
 SORT ON RATIO/D,BGFREQ/D,DESCRIPTOR TO FINAL
 USE FINAL
 ************************************************************************************************
 set talk off
 DO CASE
 CASE PTF=0
 SET DEVICE TO PRINT
 SET PRINT ON
 EJECT
 CASE PTF=1
 SET ALTERNATE TO 'Adenoid Patent Figures:Subtraction.txt'
```

```
SET ALTERNATE ON
ENDCASE

STORE VAL(SYS(2)) TO FINTIME
IF FINTIME<STARTIME
STORE FINTIME+86400 TO FINTIME
ENDIF
STORE FINTIME - STARTIME TO COMPSEC
STORE COMPSEC/60 TO COMPMIN

**********************
SET MARGIN TO 10
@1,1 SAY 'Library Subtraction Analysis' STYLE 65536 FONT 'Geneva',274 COLOR 0,0,0,-1,-1,-1
?
?
?
?
? date()
?? ' '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,5,0)
?? ' through ' '
?? STR(TERMINATE,6,0)
? 'Libraries: '
? Target1
IF Target2<>'
?? ', '
?? Target2
ENDIF
IF Target3<>'
?? ', '
?? Target3
ENDIF
? 'Subtracting:
? Object1
IF Object2<>'
?? ', '
?? Object2
ENDIF
IF Object3<>'
?? ', '
?? Object3
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF Imatch=1
?? 'INCYTE'
ENDIF
IF ANAL=1
? 'Sorted by ABUNDANCE'
ENDIF
IF ANAL=2
? 'Arranged by FUNCTION'
ENDIF
```

```
? 'Total clones represented: '
?? STR(TOT,5,0)
? 'Total clones analyzed: '
?? STR(STARTOT,5,0)
? 'Total computation time:
?? STR(COMPMIN,5,2)
?? ' minutes' '
?
? 'd = designation   f = distribution   z = location   r = function   s = species   i = inte
?
**************************
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',9 COLOR 0,0,0,
DO CASE
CASE ANAL=1
?? STR(AUNIQUE,4,0)
?? ' genes, for a total of '
?? STR(ANAL/TOT,4,0)
?? ' clones'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I
SET PRINT OFF
CLOSE DATABASES
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'

CASE ANAL=2
* arrange/function
SET PRINT ON
SET HEADING ON
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',268 COLOR 0
?
? '                              BINDING PROTEINS'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Surface molecules and receptors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='B'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Calcium-binding protsins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='C'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Ligands and effectors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='S'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Other binding protsins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='I'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',268 COLOR 0
? '                              ONCOGENES'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'General oncogenes:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='O'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'GTP-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='G'
```

```
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Viral elements:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='V'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Kinases and Phosphatases:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Y'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Tumor-related antigens:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='A'
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",268 COLOR 0
? '              PROTEIN SYNTHETIC MACHINERY PROTEINS'
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Transcription and Nucleic Acid-binding proteins:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='D'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Translation:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='T'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Ribosomal proteins:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='R'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Protein processing:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='L'
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",268 COLOR 0
?
? '                              ENZYMES'
?

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Ferroproteins:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='F'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Proteases and inhibitors:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='P'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Oxidative phosphorylation:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Z'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Sugar metabolism:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Q'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Amino acid metabolism:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
```

```
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='M'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Nucleic acid metabolism:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='N'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Lipid metabolism:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='W'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Other enzymes:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='E'
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",268 COLOR 0
?
? '                          MISCELLANEOUS CATEGORIES'
?

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Stress response:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='H'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Structural:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='K'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Other clones:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='X'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Clones of unknown function:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='U'
ENDCASE

DO "Test print.prg"
SET PRINT OFF
SET DEVICE TO SCREEN
CLOSE DATABASES
ERASE TEMPLIB.DBF
ERASE TEMPNUM.DBF
ERASE TEMPDESIG.DBF
SET MARGIN TO 0
CLEAR
LOOP
ENDDO
```

```
*Northern (single), version 11-25-94
close databases
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE '            ' TO Eobject
STORE '                              ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date....:  8/ 8/94
* Version.:.FoxBASE+/Mac, revision 1.10
* Notes....: Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 118,98 SAY 'Entry #:' STYLE 65536 FONT 'Geneva',12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT 'Geneva',12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY 'Description' STYLE 65536 FONT 'Geneva',12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT 'Geneva',12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY 'Single Northern search screen' STYLE 65536 FONT 'Geneva',274 COLOR 0,0,-1,-1,-1
@ PIXELS 220,162 GET Bail STYLE 65536 FONT 'Chicago',12 PICTURE '@*R Continue;Bail out' SIZE
@ PIXELS 175,98 SAY 'Clone #:' STYLE 65536 FONT 'Geneva',12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT 'Geneva',12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY 'Enter any ONE of the following:' STYLE 65536 FONT 'Geneva',12 COLOR -1,
*
* EOF: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE 'SmartGuy:FoxBASE+/Mac:Fox files:Lookup.dbf'
SET TALK ON

IF Eobject<>'
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO 'Lookup entry.dbf'
SET SAFETY ON
USE 'Lookup entry.dbf'
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE 'Lookup entry.dbf'
ENDIF

IF Dobject<>'
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO 'Lookup descriptor.dbf'
SET SAFETY On
USE 'Lookup descriptor.dbf'
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR
```

```
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE 'Lookup descriptor.dbf'
SET EXACT ON
ENDIF

IF Numb<>0
USE 'SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf'
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK)<>'Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE 'SmartGuy:FoxBASE+/Mac:Fox files:libraries.dbf'
SET SAFETY OFF
SORT ON library TO 'Compressed libraries.dbf'
* FOR entered>0
SET SAFETY ON
USE 'Compressed libraries.dbf'
DELETE FOR entered=0
PACK
COUNT TO TOT
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
   SW2=1
   LOOP
   ENDIF
GO MARK1
STORE library TO TESTA
SKIP
STORE Library TO TESTB
IF TESTA = TESTB
DELETE
ENDIF
MARK1 = MARK1+1
LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...'
SET TALK ON
USE 'SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf'
SET SAFETY OFF
COPY TO 'Hits.dbf' FOR entry=searchval
SET SAFETY ON
```

```
* MASTER ANALYSIS 3; VERSION 12-9-94
* Master menu for analysis output
CLOSE DATABASES
SET TALK OFF
SET SAFETY OFF
CLEAR
SET DEVICE TO SCREEN
SET DEFAULT TO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:"
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE 0 TO ENTIRE
STORE 0 TO CONDEN
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO XMATCH
STORE 0 TO PRINTON
STORE 0 TO PTF
DO WHILE .T.
* Program.: Master analysis.fmt
* Date....: 12/ 9/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Master analysis
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",9 COLOR 0,0,0,
@ PIXELS 39,255 TO 277,430 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 27,98 SAY "Customized Output Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1
@ PIXELS 45,54 GET conden STYLE 65536 FONT "Chicago",12 PICTURE "@*C Condensed format" SIZE
@ PIXELS 54,261 GET anal STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Sort/number:Sort/entry;
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Exact " SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Homologous" SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Other spc" SIZE 15,84
@ PIXELS 90,152 SAY "Matches:" STYLE 65536 FONT "Geneva",268 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 63,54 GET PRINTON STYLE 65536 FONT "Chicago",12 PICTURE "@*C Include clone listing"
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "@*C Incyte" SIZE 15,65 CO
@ PIXELS 252,146 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,146 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 234,134 SAY "Include clones " STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,125 SAY "->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",12 PICTURE "@*C Print to file" SIZE 15,9
@ PIXELS 189,0 TO 257,120 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 209,8 SAY "Library selection" STYLE 65536 FONT "Geneva",265 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 227,18 GET ENTIRE STYLE 65536 FONT "Chicago",12 PICTURE "@*RV All;Selected" SIZE 16
*
* EOF: Master analysis.fmt
READ
  IF ANAL=9
  CLEAR
  CLOSE DATABASES
  ERASE TEMPMASTER.DBF
  USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
  SET SAFETY ON
  SCREEN 1 OFF
  RETURN
  ENDIF
clear
? INITIATE
? TERMINATE
? CONDEN
? ANAL
```

```
? ematch
? Hmatch
? Omatch
? IMATCH
SET TALK ON
  IF ENTIRE=2
USE "Unique libraries.dbf"
  REPLACE ALL i WITH ' '
  BROWSE FIELDS i,libname,library,total,entered AT 0,0
  ENDIF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
*COPY TO TEMPNUM FOR NUMBER>=INITIATE.AND.NUMBER<=TERMINATE
*USE TEMPNUM
COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
  IF ENTIRE=1
  APPEND FROM 'SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf'
  ENDIF
  IF ENTIRE=2
USE 'Unique libraries.dbf'
  COPY TO SELECTED FOR UPPER(i)='Y'
  USE SELECTED
  STORE RECCOUNT() TO STOPIT
  MARK=1
    DO WHILE .T.
    IF MARK>STOPIT
    CLEAR
    EXIT
    ENDIF
    USE SELECTED
    GO MARK
    STORE library TO THISONE
    ? 'COPYING '
    ?? THISONE
    USE TEMPLIB
    APPEND FROM 'SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf' FOR library=THISONE
    STORE MARK+1 TO MARK
    LOOP
    ENDDO
  ENDIF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
COUNT TO STARTOT
COPY STRUCTURE TO TEMPDESIG
USE TEMPDESIG
  IF Ematch=0 .AND.. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
  APPEND FROM TEMPLIB
  ENDIF
  IF Ematch=1
  APPEND FROM TEMPLIB FOR D='E'
  ENDIF
  IF Hmatch=1
  APPEND FROM TEMPLIB FOR D='H'
  ENDIF
  IF Omatch=1
  APPEND FROM TEMPLIB FOR D='O'
  ENDIF
  IF Imatch=1
  APPEND FROM TEMPLIB FOR D='I'.OR.D='X'.OR.D='N'
  ENDIF
  IF Xmatch=1
  APPEND FROM TEMPLIB FOR D='X'
  ENDIF
COUNT TO ANALTOT
set talk off
**********************
DO CASE
```

```
CASE PTF=0
SET DEVICE TO PRINT
SET PRINT ON
EJECT
CASE PTF=1
SET ALTERNATE TO "Total function sort.txt"
*SET ALTERNATE TO "H and O function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance con.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Distribution sort.txt"
*SET ALTERNATE TO "Shear stress HUVEC 1:Clone list.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Location sort.txt"
SET ALTERNATE ON
ENDCASE
**********************
IF PRINTON=1
@1,30 SAY "Database Subset Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
ENDIF
?
?
?
?
? date()
?? '  '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
    MARK=1
    DO WHILE .T.
    IF MARK>STOPIT
    EXIT
    ENDIF
    USE SELECTED
    GO MARK
    ? '  '
    ?? TRIM(libname)
    STORE MARK+1 TO MARK
    LOOP
    ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF Imatch=1
?? 'INCYTE'
ENDIF
IF Xmatch=1
?? 'EST'
```

```
ENDIF
IF CONDEN=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
? 'l = library    d = designation    f = distribution    z = location    r = function    c = cer
?
*****************************************************************************************
USE TEMPDESIG
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
DO CASE
CASE ANAL=1
* sort/number
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER
DO "COMPRESSION number.PRG"
ELSE
SORT TO TEMP1 ON NUMBER
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR
*list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=2
* sort/DESCRIPTOR
SET HEADING ON
*SORT TO TEMP1 ON DESCRIPTOR,ENTRY,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
*SORT TO TEMP1 ON ENTRY,DESCRIPTOR,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
SORT TO TEMP1 ON ENTRY,START/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
IF CONDEN=1
DO "COMPRESSION entry.PRG"
ELSE
USE TEMP1
 list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF
```

```
CASE ANAL=3
* sort by abundance
SET HEADING ON
SORT TO TEMP1 ON ENTRY,NUMBER for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
DO "COMPRESSION abundance.PRG"

CASE ANAL=4
* sort/interest
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER FOR I>0
DO "COMPRESSION interest.PRG"
ELSE
SORT ON I/D,ENTRY TO TEMP1 FOR I>1
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=5
* arrange/location
SET HEADING ON
STORE 4 TO AMPLIFIER
? 'Nuclear:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoplasmic:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoskeleton:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cell surface:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Intracellular membrane:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Mitochondrial:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
```

```
? 'Secreted:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDEN=1
SET DEVICE TO PRINTER
SET PRINTER ON
EJECT
DO "Output heading.prg"
USE "Analysis location.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
? '      FUNCTIONAL CLASS                    TOTAL   UNIQUE  NEW  % TOTAL'
?
LIST OFF FIELDS Z,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF

CASE ANAL=6
* arrange/distribution
SET HEADING ON
STORE 3 TO AMPLIFIER
? 'Cell/tissue specific distribution:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Non-specific distribution:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown distribution:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDEN=1
SET DEVICE TO PRINTER
SET PRINTER ON
```

```
EJECT
DO "Output heading.prg"
USE "Analysis distribution.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
? '      FUNCTIONAL CLASS                    TOTAL   UNIQUE  % TOTAL'
?
LIST OFF FIELDS P.NAME,CLONES,GENES,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE 'SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF

CASE ANAL=7
* arrange/function
SET HEADING ON
STORE 10 TO AMPLIFIER
? '                              BINDING PROTEINS'
?
? 'Surface molecules and receptors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Calcium-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Ligands and effectors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                              ONCOGENES'
?
? 'General oncogenes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'GTP-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Viral elements:'
```

```
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Kinases and Phosphatases:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Tumor-related antigens:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
'EJECT
? '                              PROTEIN SYNTHETIC MACHINERY PROTEINS'
?
? 'Transcription and Nucleic Acid-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Translation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Ribosomal proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Protein processing:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
'EJECT
? '                                    ENZYMES'
?
? 'Ferroproteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Proteases and inhibitors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
```

```
DO "Normal subroutine 1"
ENDIF
? 'Oxidative phosphorylation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Sugar metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Amino acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Nucleic acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Lipid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other enzymes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                                MISCELLANEOUS CATEGORIES'
?
? 'Stress response:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Structural:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other clones:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
```

```
DO 'Normal subroutine 1'
ENDIF
? 'Clones of unknown function:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF

IF CONDEN=1
EJECT
*SET DEVICE TO PRINTER
*SET PRINT ON
DO 'Output heading.prg'
***
USE 'Analysis function.dbf'
DO 'Create bargraph.prg'
SET HEADING OFF
***
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,0
***
? '
? '          FUNCTIONAL CLASS                              CLONES   GENES  TOTAL GENES  TOTAL NEW FUNCTIONAL CLASS'  DIST
? '
***
*LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH,COMPANY
LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE 'SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf'
ENDIF
CASE ANAL=8
DO "Subgroup summary 3.prg"
ENDCASE
DO "Test print.prg"
SET PRINT OFF
SET DEVICE TO SCREEN
CLOSE DATABASES
*ERASE TEMPLIB.DBF
*ERASE TEMPNUM.DBF
*ERASE TEMPDESIG.DBF
*ERASE SELECTED.DBF
CLEAR
LOOP
ENDDO
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  COUNT TO NEWGENES FOR D='H'.OR.D='O'
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE Z TO LOC
USE "Analysis location.dbf"
LOCATE FOR Z=LOC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                          V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON DATE TO TEMP2
USE TEMP2
?? STR(UNIQUE,4,0)
?? ' genes, for a total of'
?? STR(TOT,4,0)
?? ' clones'
?
? '                          V Coincidence'
COUNT TO P4 FOR I=4
IF P4>0
? STR(P4,3,0)
?? ' genes with priority = 4 (Secondary analysis:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=4
?
ENDIF
COUNT TO P3 FOR I=3
IF P3>0
? STR(P3,3,0)
?? ' genes with priority = 3 (Full insert sequence:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=3
?
ENDIF
COUNT TO P2 FOR I=2
IF P2>0
? STR(P2,3,0)
?? ' genes with priority = 2 (Primary analysis complete:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=2
?
ENDIF
COUNT TO P1 FOR I=1
IF P1>0
```

```
? STR(P1,3,0)
?? ' genes with priority = 1 (Primary analysis needed:)'
list off fields number,RFEND,L,D,P,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=1
ENDIF

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON NUMBER TO TEMP2
USE TEMP2

?? STR(UNIQUE,4,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
?                          V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  COUNT TO NEWGENES FOR D='H'.OR.D='O'
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE R TO FUNC
USE "Analysis function.dbf"
LOCATE FOR P=FUNC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
SET HEADING ON
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
***
? '                    V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I
***
*SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,
*list off fields RFEND,S,DESCRIPTOR

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE F TO DIST
USE "Analysis distribution.dbf"
LOCATE FOR P=DIST
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
USE TEMP1
sort on rfend/d to TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                    V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
USE TEMP1
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                              V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
COPY TO TEMP1 FOR
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '  Coincidence V       V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",9 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '  Coincidence V       V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
*Lifescan menu; version 8-7-94
SET TALK OFF
set device to screen
CLEAR
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
STORE LUPDATE() TO Update
GO BOTTOM
STORE RECNO() TO cloneno
STORE 6 TO Chooser
DO WHILE .T.
* Program.: Lifeseq menu.fmt
* Date....:  1/11/95
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Lifeseq menu
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",268 COLOR 0,0,
@ PIXELS 18,126 TO 77,365 STYLE 28479 COLOR 32767,-25600,-1,-16223,-16721,-15725
@ PIXELS 110,29 TO 188,217 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 45,161 SAY "LIFESEQ" STYLE 65536 FONT "Geneva",536 COLOR 0,0,-1,-1,7135,5884
@ PIXELS 36,269 SAY "TM" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,7135,5884
@ PIXELS 63,143 SAY "Molecular Biology Desktop" STYLE 65536 FONT "Helvetica",18 COLOR 0,0,0,
@ PIXELS 90,252 TO 251,467 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 117,270 GET Chooser STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Transcript profiles
@ PIXELS 135,128 SAY Update STYLE 0 FONT "Geneva",12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 171,128 SAY cloneno STYLE 0 FONT "Geneva",12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 135,44 SAY "Last update:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 171,44 SAY "Total clones:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 45,296 SAY "v1.30" STYLE 65536 FONT "Geneva",782 COLOR 0,0,-1,-1,-1,-1
*
* EOF: Lifeseq menu.fmt
READ
DO CASE
CASE Chooser=1
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Master analysis 3.prg"
CASE Chooser=2
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Subtraction 2.prg"
CASE Chooser=3
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Northern (single).prg"
CASE Chooser=4
USE "Libraries.dbf"
BROWSE
CASE Chooser=5
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:See individual clone.prg"
CASE Chooser=6
DO "SmartGuy:FoxBASE+/Mac:fox files:Libraries:Output programs:Menu.prg"
CASE Chooser=7
CLEAR
SCREEN 1 OFF
RETURN
ENDCASE

LOOP
ENDDO
```

```
@1,30 SAY "Database Subset Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
?
?
?
?
? date()
?? '  '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
    MARK=1
    DO WHILE .T.
    IF MARK>STOPIT
    EXIT
    ENDIF
    USE SELECTED
    GO MARK
    ? '  '
    ?? TRIM(libname)
    STORE MARK+1 TO MARK
    LOOP
    ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF CONDEN=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
?'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'
```

```
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
?
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones!'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
*Northern (single), version 11-25-94
close databases
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE '          ' TO Eobject
STORE '                                ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date....:  8/ 8/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,0
@ PIXELS 15,81 TO 45,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY "Entry #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT "Geneva",12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY "Description" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT "Geneva",12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY "Single Northern search screen" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Continue;Bail out" SIZE
@ PIXELS 175,98 SAY "Clone #:" STYLE 65536 FONT "Geneva";12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY "Enter any ONE of the following:" STYLE 65536 FONT "Geneva",12 COLOR -1,
*
* EOF: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE "SmartGuy:FoxBASE+/Mac:Fox files:Lookup.dbf"
SET TALK ON

IF Eobject<>'          '
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO "Lookup entry.dbf"
SET SAFETY ON
USE "Lookup entry.dbf"
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup entry.dbf"
ENDIF

IF Dobject<>'                                '
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO "Lookup descriptor.dbf"
SET SAFETY On
USE "Lookup descriptor.dbf"
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR
```

```
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup descriptor.dbf"
SET EXACT ON
ENDIF

IF Numb<>0
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK)<>'Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE "SmartGuy:FoxBASE+/Mac:Fox files:libraries.dbf"
SET SAFETY OFF
SORT ON library TO "Compressed libraries.dbf"
* FOR entered>0
SET SAFETY ON
USE "Compressed libraries.dbf"
DELETE FOR entered=0
PACK
COUNT TO TOT
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  SW2=1
  LOOP
  ENDIF
GO MARK1
STORE library TO TESTA
SKIP
STORE Library TO TESTB
IF TESTA = TESTB
DELETE
ENDIF
MARK1 = MARK1+1
LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...'
SET TALK ON
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
SET SAFETY OFF
COPY TO "Hits.dbf" FOR entry=searchval
SET SAFETY ON
```

```
CLOSE DATABASES
SELECT 1
USE "Compressed libraries.dbf"
STORE RECCOUNT() TO Entries
SELECT 2
USE "Hits.dbf"
Mark=1
DO WHILE .T.
SELECT 1
IF Mark>Entries
EXIT
ENDIF
GO MARK
STORE library TO Jigger
SELECT 2
COUNT TO Zog FOR library=Jigger
SELECT 1
REPLACE hits with Zog
Mark=Mark+1
LOOP
ENDDO

SELECT 1
BROWSE FIELDS LIBRARY,LIBNAME,ENTERED,HITS AT 0,0
CLEAR
? 'Enter Y to print:'
WAIT TO PRINSET
IF UPPER(PRINSET)='Y'
SET PRINT ON
CLEAR
EJECT
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",14 COLOR 0,0,0
? 'DATABASE ENTRIES MATCHING ENTRY '
?? Searchval
? DATE()
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
LIST OFF FIELDS library,libname,entered,hits
?
?
SELECT 2
LIST OFF FIELDS NUMBER,LIBRARY,D,S,F,Z,R,ENTRY,DESCRIPTOR,RFSTART,START,RFEND
SET TALK OFF
SET PRINT OFF
ENDIF
CLOSE DATABASES
SET TALK OFF
CLEAR
DO "Test print.prg"
RETURN
```

TABLE 6

| library | libname |
|---|---|
| ADENINB01 | Inflamed adenoid |
| ADRENOR01 | Adrenal gland (r) |
| ADRENOT01 | Adrenal gland (T) |
| AMLBNOT01 | AML blast cells (T) |
| BMARNOT01 | Bone marrow |
| BMARNOT02 | Bone marrow (T) |
| CARDNOT01 | Cardiac muscle (T) |
| CHAONOT01 | Chin. hamster ovary |
| CORNNOT01 | Corneal stroma |
| FIBRAGT01 | Fibroblast, AT 5 |
| FIBRAGT02 | Fibroblast, AT 30 |
| FIBRANT01 | Fibroblast, AT |
| FIBRNGT01 | Fibroblast, uv 5 |
| FIBRNGT02 | Fibroblast, uv 30 |
| FIBRNOT01 | Fibroblast |
| FIBRNOT02 | Fibroblast, normal |
| HMC1NOT01 | Mast cell line HMC-1 |
| HUVELPB01 | HUVEC IFN,TNF,LPS |
| HUVENOB01 | HUVEC control |
| HUVESTB01 | HUVEC shear stress |
| HYPONOB01 | Hypothalamus |
| KIDNNOT01 | Kidney (T) |
| LIVRNOT01 | Liver (T) |
| LUNGNOT01 | Lung (T) |
| MUSCNOT01 | Skeletal muscle (T) |
| OVIDNOB01 | Oviduct |
| PANCNOT01 | Pancreas, normal |
| PITUNOR01 | Pituitary (r) |
| PITUNOT01 | Pituitary (T) |
| PLACNOB01 | Placenta |
| SINTNOT02 | Small intestine (T) |
| SPLNFET01 | Spleen+liver, fetal |
| SPLNNOT02 | Spleen (T) |
| STOMNOT01 | Stomach |
| SYNORAB01 | Rheum. synovium |
| TBLYNOT01 | T + B lymphoblast |
| TESTNOT01 | Testis (T) |
| THP1NOB01 | THP-1 control |
| THP1PEB01 | THP phorbol |
| THP1PLB01 | THP-1 phorbol LPS |
| U937NOT01 | U937, monocytic leuk |

| number | library | d | s | f | z | r | entry | descriptor | rfstart | rfend |
|---|---|---|---|---|---|---|---|---|---|---|
| 2304 | U937NOT01 | E | H | C | C | T | HUMEF1B | Elongation factor 1-beta | 0 | 773 |
| 3240 | HMC1NOT01 | E | H | C | C | T | HUMEF1B | Elongation factor 1-beta | 0 | 773 |
| 3259 | HMC1NOT01 | E | H | C | C | T | HUMEF1B | Elongation factor 1-beta | 0 | 773 |
| 4693 | HMC1NOT01 | E | H | C | C | T | HUMEF1B | Elongation factor 1-beta | 0 | 773 |
| 8989 | HMC1NOT01 | E | H | C | C | T | HUMEF1B | Elongation factor 1-beta | 0 | 773 |
| 9139 | HMC1NOT01 | E | H | C | C | T | HUMEF1B | Elongation factor 1-beta | 0 | 773 |

WHAT IS CLAIMED IS:

1. A method of analyzing a specimen containing gene transcripts, said method comprising the steps of:

(a) producing a library of biological sequences;

(b) generating a set of transcript sequences, where each of the transcript sequences in said set is indicative of a different one of the biological sequences of the library;

(c) processing the transcript sequences in a programmed computer in which a database of reference transcript sequences indicative of reference biological sequences is stored, to generate an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and a degree of match between one of the transcript sequences and at least one of the reference transcript sequences; and

(d) processing each said identified sequence value to generate final data values indicative of a number of times each identified sequence value is present in the library.

2. The method of claim 1, wherein step (a) includes the steps of:

obtaining a mixture of mRNA;

making cDNA copies of the mRNA;

isolating a representative population of clones transfected with the cDNA and producing therefrom the library of biological sequences.

3. The method of claim 1, wherein the biological sequences are cDNA sequences.

4. The method of claim 1, wherein the biological sequences are RNA sequences.

5. The method of claim 1, wherein the biological sequences are protein sequences.

6. The method of claim 1, wherein a first value of said degree of match is indicative of an exact match, and a second value of said degree of match is indicative of a non-exact match.

7. A method of comparing two specimens containing gene transcripts, said method comprising:

(a) analyzing a first specimen according to the method of claim 1;

(b) producing a second library of biological sequences;

(c) generating a second set of transcript sequences, where each of the transcript sequences in said second set is indicative of a different one of the biological sequences of the second library;

(d) processing the second set of transcript sequences in said programmed computer to generate a second set of identified sequence values known as further identified sequence values, where each of the further identified sequence values is indicative of a sequence annotation and a degree of match between one of the biological sequences of the second library and at least one of the reference sequences;

(e) processing each said further identified sequence value to generate further final data values indicative of a number of times each further identified sequence value is present in the second library; and

(f) processing the final data values from the first specimen and the further identified sequence values from the second specimen to generate ratios of transcript sequences, each of said ratio values indicative of differences in numbers of gene transcripts between the two specimens.

8. A method of quantifying relative abundance of mRNA in a biological specimen, said method comprising the steps of:

(a) isolating a population of mRNA transcripts from the biological specimen;

(b) identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c) determining numbers of mRNA transcripts corresponding to each of the genes; and

(d) using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts.

9. A diagnostic method which comprises producing a gene transcript image, said method comprising the steps of:

(a) isolating a population of mRNA transcripts from a biological specimen;

(b) identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c) determining numbers of mRNA transcripts corresponding to each of the genes; and

(d) using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts, where data determining the relative abundance values of mRNA transcripts is the gene transcript image of the biological specimen.

10. The method of claim 9, further comprising:

(e) providing a set of standard normal and diseased gene transcript images; and

(f) comparing the gene transcript image of the biological specimen with the gene transcript images of step (e) to identify at least one of the standard gene transcript images which most closely approximate the gene transcript image of the biological specimen.

11. The method of claim 9, wherein the biological specimen is biopsy tissue, sputum, blood or urine.

12. A method of producing a gene transcript image, said method comprising the steps of

(a) obtaining a mixture of mRNA;

(b) making cDNA copies of the mRNA;

(c) inserting the cDNA into a suitable vector and using said vector to transfect suitable host strain cells which are plated out and permitted to grow into clones, each clone representing a unique mRNA;

(d) isolating a representative population of recombinant clones;

(e) identifying amplified cDNAs from each clone in the population by a sequence-specific method which identifies gene from which the unique mRNA was transcribed;

(f) determining a number of times each gene is represented within the population of clones as an indication of relative abundance; and

(g) listing the genes and their relative abundance in order of abundance, thereby producing the gene transcript image.

13. The method of claim 12, also including the step of diagnosing disease by:

repeating steps (a) through (g) on biological specimens from random sample of normal and diseased humans, encompassing a variety of diseases, to produce reference sets of normal and diseased gene transcript images;

obtaining a test specimen from a human, and producing a test gene transcript image by performing steps (a) through (g) on said test specimen;

comparing the test gene transcript image with the reference sets of gene transcript images; and

identifying at least one of the reference gene transcript images which most closely approximates the test gene transcript image.

14. A computer system for analyzing a library of biological sequences, said system including:

means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a different one of the biological sequences of the library; and

means for processing the transcript sequences in the computer system in which a database of reference transcript

sequences indicative of reference biological sequences is stored, wherein the computer is programmed with software for generating an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and a degree of match between a different one of the biological sequences of the library and at least one of the reference transcript sequences, and for processing each said identified sequence value to generate final data values indicative of a number of times each identified sequence value is present in the library.

15. The system of claim 14, also including:

library generation means for producing the library of biological sequences and generating said set of transcript sequences from said library.

16. The system of claim 15, wherein the library generation means includes:

means for obtaining a mixture of mRNA;

means for making cDNA copies of the mRNA;

means for inserting the cDNA copies into cells and permitting the cells to grow into clones;

means for isolating a representative population of the clones and producing therefrom the library of biological sequences.

# SYBASE database Structure

## Library Preparation

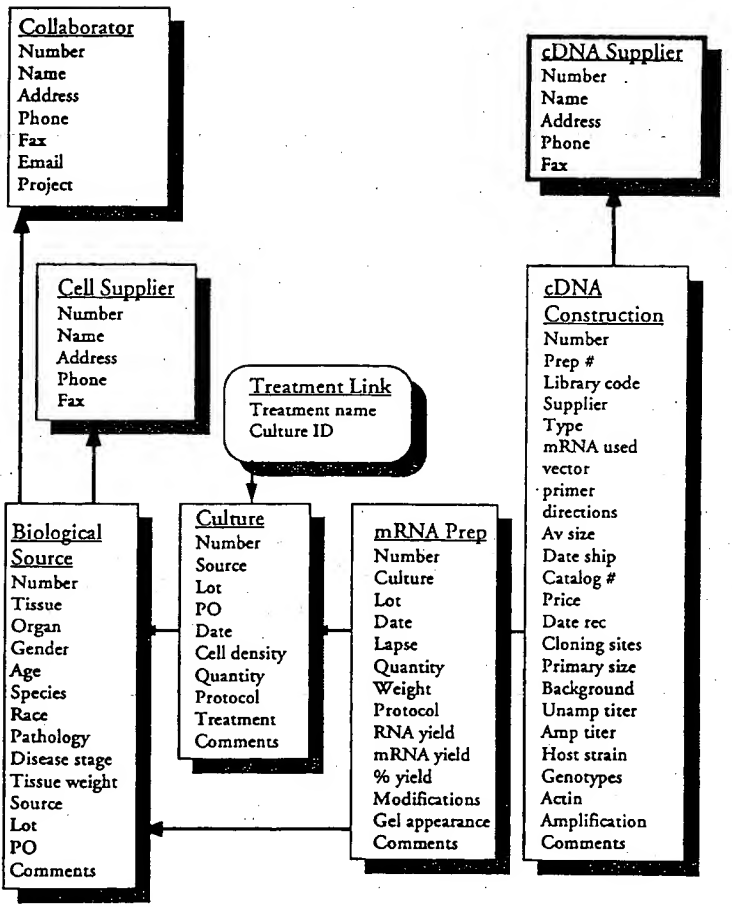



Figure 1

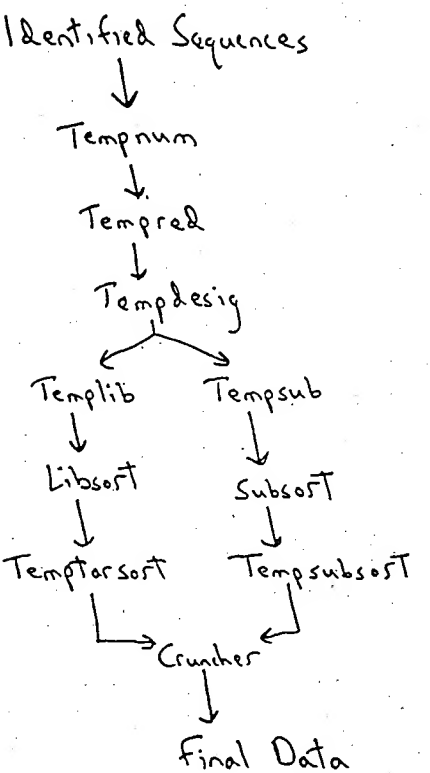

Identified Sequences
Tempnum
Tempred
Tempdesig
Templib
Tempsub
Libsort
Subsort
Temptarsort
Tempsubsort
Cruncher
Final Data

Figure 2

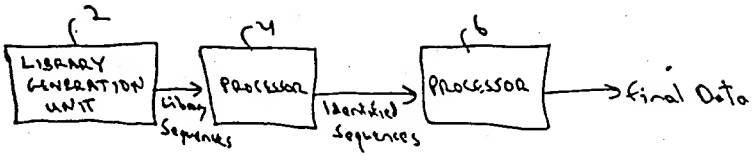

2
LIBRARY GENERATION UNIT
Library Sequences
4
PROCESSOR
Identified Sequences
6
PROCESSOR
Final Data

Figure 3

# Incyte Bioinformatics Process

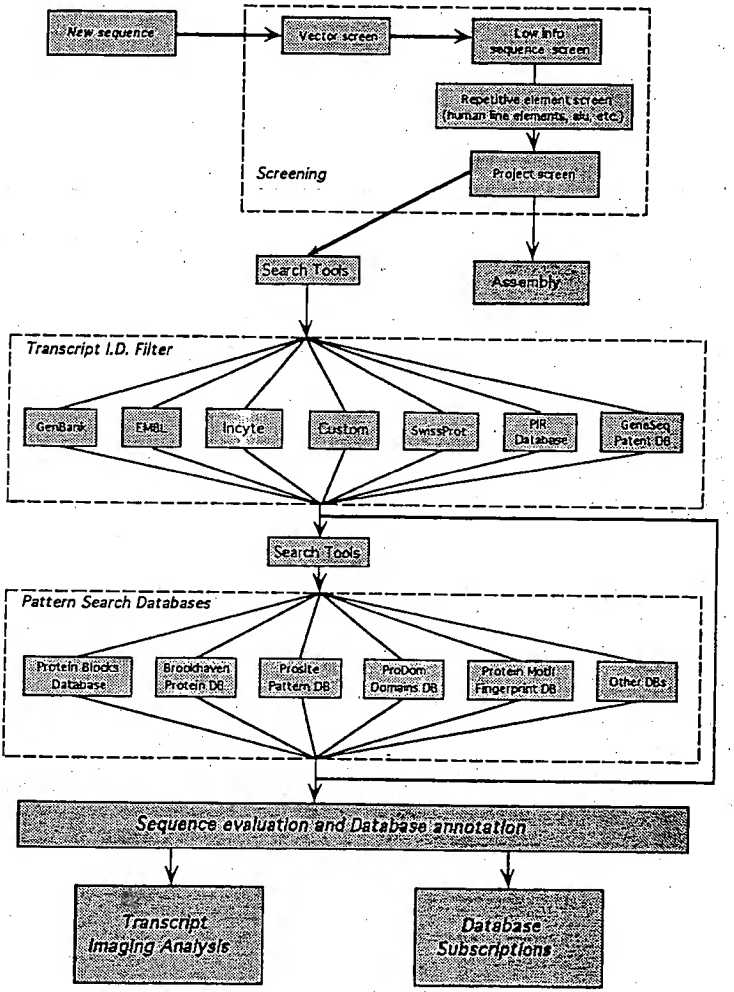



Figure 4

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01160

**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(6) :C12Q 1/68; G06F 15/00
US CL : 435/6; 364/413.02
According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 364/413.02

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CAS ONLINE, APS, transcript, transcripts, cdan#, mrna#, frequenc?, distribut?, abundanc?

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X<br>---<br>Y | IntelliGenetics Suite, Release 5.4, Advanced Training Manual, issued January 1993 by IntelliGenetics, Inc. 700 East El Camino Real, Mountain View, California 94040, United States of America, pages (1-6)-(1-19) and (2-9)-(2-14), see entire document. | 15 and 16<br>----------<br>1-14 |
| Y | Science, Volume 252, issued 21 June 1991, M.D. Adams et al, "Complementary DNA sequencing: Expressed sequence tags and human genome project", pages 1651-1656, see entire document. | 1-16 |

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

| | | | |
|---|---|---|---|
| * | Special categories of cited documents: | "T" | later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention |
| "A" | document defining the general state of the art which is not considered to be of particular relevance | "X" | document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone |
| "E" | earlier document published on or after the international filing date | "Y" | document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art |
| "L" | document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) | "&" | document member of the same patent family |
| "O" | document referring to an oral disclosure, use, exhibition or other means | | |
| "P" | document published prior to the international filing date but later than the priority date claimed | | |

| | |
|---|---|
| Date of the actual completion of the international search<br>27 APRIL 1995 | Date of mailing of the international search report<br>04 MAY 1995 |
| Name and mailing address of the ISA/US<br>Commissioner of Patents and Trademarks<br>Box PCT<br>Washington, D.C. 20231<br>Facsimile No. (703) 305-3230 | Authorized officer<br>JAMES MARTINELL<br>Telephone No. (703) 308-0196 |

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/01160

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y | Nucleic Acids Research, Volume 19, No. 25, issued 1991, E. Hara et al, "Subtractive cDNA cloning using oligo($dT)_{30}$-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells", pages 7097-7104, see entire document. | 1-16 |
| X<br>---<br>Y | Nature Genetics, Volume 2, No. 3, issued November 1992, K. Okubo et al, "Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression", pages 173-179, see narrative text portion of entire document. | 1, 3<br>---<br>2 and 4-16 |

Ax1p sequence following $Ser^{206}$ and occurs within the domain of Ax1p that shows homology with hIDE (*14*). To delete the complete *STE23* sequence and create the *ste23Δ::URA3* mutation, polymerase chain reaction (PCR) primers (5′-TCGGAAGACCTCAT-TCTTGCTCATTTTGATATTGCTC- TGTAGATTG-TACTGAGAGTGCAC-3′; and 5′-GCTACAAACAGC-GTCGACTTGAATGCCCCGACATCTTCGACTGT-GCGGTATTTCACACCG-3′) were used to amplify the *URA3* sequence of pRS316, and the reaction product was transformed into yeast for one-step gene replacement [R. Rothstein, *Methods Enzymol.* 194, 281 (1991)]. To create the *axl1Δ::LEU2* mutation contained on p114, a 5.0-kb Sal I fragment from p*AXL1* was cloned into pUC19, and an internal 4.0-kb Hpa I–Xho I fragment was replaced with a *LEU2* fragment. To construct the *ste23Δ::LEU2* allele (a deletion corresponding to 931 amino acids) carried on p153, a *LEU2* fragment was used to replace the 2.8-kb Pml I–Ecl136 II fragment of *STE23*, which occurs within a 6.2-kb Hind III–Bgl II genomic fragment carried on pSP72 (Promega). To create Y*EpMFA1*, a 1.5-kb Bam HI fragment containing *MFA1*, from pKK16 [K. Kuchler, H. E. Sterne, J. Thorner, *EMBO J.* 8, 3973 (1989)], was ligated into the Bam HI site of YEp351 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)].

24. J. Chant and I. Herskowitz, *Cell* 65, 1203 (1991).
25. B. W. Matthews, *Acc. Chem. Res.* 21, 333 (1988).
26. K. Kuchler, H. G. Dohlman, J. Thorner; *J. Cell Biol.* 120, 1203 (1993); R. Kolling and C. P. Hollenberg, *EMBO J.* 13, 3261 (1994); C. Berkower, D. Loayza, S. Michaelis, *Mol. Biol. Cell* 5, 1185 (1994).
27. A. Bender and J. R. Pringle, *Proc. Natl. Acad. Sci. U.S.A* 86, 9976 (1989); J. Chant, K. Corrado, J. R. Pringle, I. Herskowitz, *Cell* 65, 1213 (1991); S. Powers, E. Gonzales, T. Christensen, J. Cubert, D. Broek, *ibid.*, p. 1225; H. O. Park, J. Chant, I. Herskowitz, *Nature* 365, 269 (1993); J. Chant, *Trends Genet.* 10, 328 (1994); _____ and J. R. Pringle, *J. Cell Biol.* 129, 751 (1995); J. Chant, M. Mischke, E. Mitchell, I. Herskowitz, J. R. Pringle, *ibid.*, p. 767.
28. G. F. Sprague Jr., *Methods. Enzymol.* 194, 77 (1991).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
30. A W303 1A derivative, SY2625 (*MATa ura3-1 leu2-3, -112 trp1-1 ade2-1 can1-100 sst1Δ mfa2Δ::FUS1-lacZ his3Δ::FUS1-HIS3*), was the parent strain for the mutant search. SY2625 derivatives for the mating assays, secreted pheromone assays, and the pulse-chase experiments included the following strains: Y49 (*ste22-1*), Y115 (*mfa1Δ::LEU2*), Y142 (*axl1::URA3*), Y173 (*axl1Δ::LEU2*), Y220 (*axl1::URA3 ste23Δ::URA3*), Y221 (*ste23Δ::URA3*), Y231 (*axl1Δ::LEU2 ste23Δ::LEU2*), and Y233 (*ste23Δ::LEU2*). *MATα* derivatives of SY2625 included the following strains: Y199 (SY2625 made *MATα*), Y278 (*ste22-1*), Y195 (*mfa1Δ::LEU2*), Y196 (*axl1Δ::LEU2*), and Y197 (*axl1::URA3*). The EG123 (*MATa leu2 ura3 trp1 can1 his4*) genetic background was used to create a set of strains for analysis of bud site selection. EG123 derivatives included the following strains: Y175 (*axl1Δ::LEU2*), Y223 (*axl1::URA3*), Y234 (*ste23Δ::LEU2*), and Y272 (*axl1Δ::LEU2 ste23Δ::LEU2*). *MATα* derivatives of EG123 included the following strains: Y214 (EG123 made *MATα*) and Y293 (*axl1Δ::LEU2*). All strains were generated by means of standard genetic or molecular methods involving the appropriate constructs (*23*). In particular, the *axl1 ste23* double mutant strains were created by crossing of the appropriate *MATa ste23* and *MATα axl1* mutants, followed by sporulation of the resultant diploid and isolation of the double mutant from nonparental di-type tetrads. Gene disruptions were confirmed with either PCR or Southern (DNA) analysis.
31. p129 is a YEp352 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, *Yeast* 2, 163 (1986)] plasmid containing a 5.5-kb Sal I fragment of p*AXL1*. p151 was derived from p129 by insertion of a linker at the Bgl II site within *AXL1*, which led to an in-frame insertion of the hemagglutinin (HA) epitope (DDYPYDVPDYA) (*29*) between amino acids 854 and 855 of the *AXL1* product. pC225 is a KS+ (Stratagene) plasmid containing a 0.5-kb Bam HI–Sst I fragment from p*AXL1*. Substitution mutations of the proposed active site of Axl1p were created with the use of pC225 and site-specific mutagenesis involving appropriate synthetic oligonucleotides (*axl1-H68A*, 5′-GTGCTCACAAAGCGCT-GCCAAACCGCC-3′; *axl1-E71A*, 5′-AAGAATCAT-GTGCGCACAAAGGTGCGC-3′; and *axl1-E71D*, 5′-AAGAATCATGTGATCACAAAGGTGCGC-3′). The mutations were confirmed by sequence analysis. After mutagenesis, the 0.4-kb Bam HI–Msc I fragment from the mutagenized pC225 plasmids was transferred into p*AXL1* to create a set of pRS316 plasmids carrying different *AXL1* alleles, p124 (*axl1-H68A*), p130 (*axl1-E71A*), and p132 (*axl1-E71D*). Similarly, a set of HA-tagged alleles carried on YEp352 were created after replacement of the p151 Bam HI–Msc I fragment, to generate p161 (*axl1-E71A*), p152 (*axl1-*

32

N. Davis, T. Favero, C. de Hoog, and S. Kim for comments on the manuscript. Supported by a grant to C.B. from the Natural Sciences and Engineering Research Council of Canada. Support for M.N.A. was from a California Tobacco-Related Disease Research Program postdoctoral fellowship (4FT-0083).

22 June 1995; accepted 21 August 1995

# Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis, Patrick O. Brown‡

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 *Arabidopsis* genes were made by means of simultaneous, two-color fluorescence hybridization.

The temporal, developmental, topographical, histological, and physiological patterns in which a gene is expressed provide clues to its biological role. The large and expanding database of complementary DNA (cDNA) sequences from many organisms (*1*) presents the opportunity of defining these patterns at the level of the whole genome.

For these studies, we used the small flowering plant *Arabidopsis thaliana* as a model organism. *Arabidopsis* possesses many advantages for gene expression analysis, including the fact that it has the smallest genome of any higher eukaryote examined to date (*2*). Forty-five cloned *Arabidopsis* cDNAs (Table 1), including 14 complete sequences and 31 expressed sequence tags (ESTs), were used as gene-specific targets. We obtained the ESTs by selecting cDNA clones at random from an *Arabidopsis* cDNA library. Sequence analysis revealed that 28 of the 31 ESTs matched sequences in the database (Table 1). Three additional cDNAs from other organisms served as controls in the experiments.

The 48 cDNAs, averaging ~1.0 kb, were amplified with the polymerase chain reaction (PCR) and deposited into individual wells of a 96-well microtiter plate. Each sample was duplicated in two adjacent wells to allow the reproducibility of the arraying and hybridization process to be tested. Samples from the microtiter plate were printed onto glass microscope slides in an area measuring 3.5 mm by 5.5 mm with the use of a high-speed arraying machine (*3*). The arrays were processed by chemical and heat treatment to attach the DNA sequences to the glass surface and denature them (*3*). Three arrays, printed in a single lot, were used for the experiments here. A single microtiter plate of PCR products provides sufficient material to print at least 500 arrays.

Fluorescent probes were prepared from total *Arabidopsis* mRNA (*4*) by a single round of reverse transcription (*5*). The *Arabidopsis* mRNA was supplemented with human acetylcholine receptor (AChR) mRNA at a dilution of 1:10,000 (w/w) before cDNA synthesis, to provide an internal standard for calibration (*5*). The resulting fluorescently labeled cDNA mixture was hybridized to an array at high stringency (*6*) and scanned

M. Schena and R. W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.
D. Shalon and P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

*These authors contributed equally to this work.
†Present address: Synteni, Palo Alto, CA 94303, USA.
‡To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

with a laser (3). A high-sensitivity scan gave signals that saturated the detector at nearly all of the *Arabidopsis* target sites (Fig. 1A). Calibration relative to the AChR mRNA standard (Fig. 1A) established a sensitivity limit of ~1:50,000. No detectable hybridization was observed to either the rat glucocorticoid receptor (Fig. 1A) or the yeast *TRP4* (Fig. 1A) targets even at the highest scanning sensitivity. A moderate-sensitivity scan of the same array allowed linear detection of the more abundant transcripts (Fig. 1B). Quantitation of both scans revealed a range of expression levels spanning three orders of magnitude for the 45 genes tested (Table 2). RNA blots (7) for several genes (Fig. 2) corroborated the expression levels measured with the microarray to within a factor of 5 (Table 2).

Differential gene expression was investigated with a simultaneous, two-color hybridization scheme, which served to minimize experimental variation inherent in the comparison of independent hybridizations. Fluorescent probes were prepared from two mRNA sources with the use of reverse transcriptase in the presence of fluorescein- and lissamine-labeled nucleotide analogs, respectively (5). The two probes were then mixed together in equal proportions, hybridized to a single array, and scanned separately for fluorescein and lissamine emission after independent excitation of the two fluorophores (3).

To test whether overexpression of a single gene could be detected in a pool of total *Arabidopsis* mRNA, we used a microarray to analyze a transgenic line overexpressing the single transcription factor *HAT4* (8). Fluorescent probes representing mRNA from wild-type and *HAT4*-transgenic plants were labeled with fluorescein and lissamine, respectively; the two probes were then mixed and hybridized to a single array. An intense hybridization signal was observed at the position of the *HAT4* cDNA in the lissamine-specific scan (Fig. 1D), but not in the fluorescein-specific scan of the same array (Fig. 1C). Calibration with AChR mRNA added to the fluorescein and lissamine cDNA synthesis reactions at dilutions of 1:10,000 (Fig. 1C) and 1:100 (Fig. 1D), respectively, revealed a 50-fold elevation of *HAT4* mRNA in the transgenic line relative to its abundance in wild-type plants (Table 2). This magnitude of *HAT4* overexpression matched that inferred from the Northern (RNA) analysis within a factor of 2 (Fig. 2 and Table 2). Expression of all the other genes monitored on the array differed by less than a factor of 5 between *HAT4*-transgenic and wild-type plants (Fig 1, C

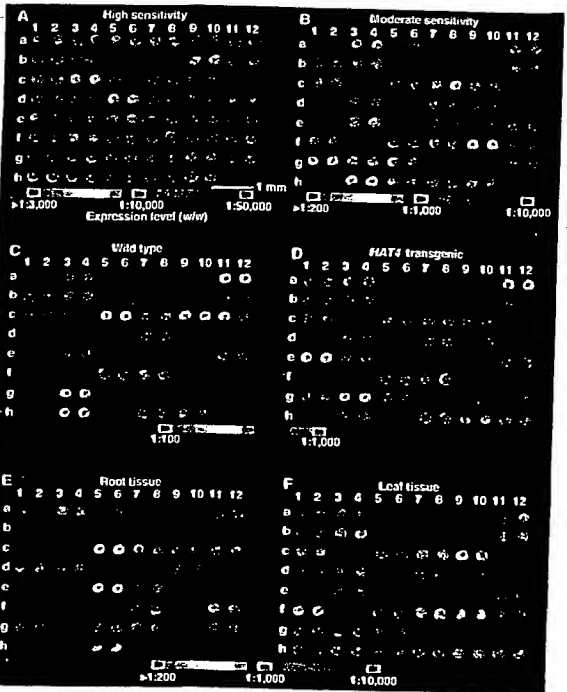



**Fig. 1.** Gene expression monitored with the use of cDNA microarrays. Fluorescent scans represented in pseudocolor correspond to hybridization intensities. Color bars were calibrated from the signal obtained with the use of known concentrations of human AChR mRNA in independent experiments. Numbers and letters on the axes mark the position of each cDNA. (A) High-sensitivity fluorescein scan after hybridization with fluorescein-labeled cDNA derived from wild-type plants. (B) Same array as in (A) but scanned at moderate sensitivity. (C and D) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from wild-type plants and lissamine-labeled cDNA from HAT4-transgenic plants. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNA from wild-type plants (C) and the lissamine fluorescence corresponding to mRNA from HAT4-transgenic plants (D). (E and F) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from root tissue and lissamine-labeled cDNA from leaf tissue. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNAs expressed in roots (E) and the lissamine fluorescence corresponding to mRNAs expressed in leaves (F).

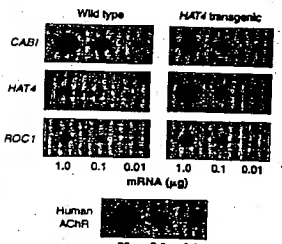



**Fig. 2.** Gene expression monitored with RNA (Northern) blot analysis. Designated amounts of mRNA from wild-type and *HAT4*-transgenic plants were spotted onto nylon membranes and probed with the cDNAs indicated. Purified human AChR mRNA was used for calibration.

and D, and Table 2). Hybridization of fluorescein-labeled glucocorticoid receptor cDNA (Fig. 1C) and lissamine-labeled *TRP4* cDNA (Fig. 1D) verified the presence of the negative control targets and the lack of optical cross talk between the two fluorophores.

To explore a more complex alteration in expression patterns, we performed a second two-color hybridization experiment with fluorescein- and lissamine-labeled probes prepared from root and leaf mRNA, respectively. The scanning sensitivities for the two fluorophores were normalized by matching the signals resulting from AChR mRNA, which was added to both cDNA synthesis reactions at a dilution of 1:1000 (Fig. 1, E and F). A comparison of the scans revealed widespread differences in gene expression between root and leaf tissue (Fig. 1, E and F). The mRNA from the light-regulated *CAB1* gene was ~500-fold more abundant in leaf (Fig. 1F) than in root tissue (Fig. 1E). The expression of 26 other genes differed between root and leaf tissue by more than a factor of 5 (Fig. 1, E and F).

The *HAT4*-transgenic line we examined has elongated hypocotyls, early flowering, poor germination, and altered pigmentation (*8*). Although changes in expression were observed for *HAT4*, large changes in expression were not observed for any of the other 44 genes we examined. This was somewhat surprising, particularly because comparative analysis of leaf and root tissue identified 27 differentially expressed genes. Analysis of an expanded set of genes may be required to identify genes whose expression changes upon *HAT4* overexpression; alternatively, a comparison of mRNA populations from specific tissues of wild-type and *HAT4*-transgenic plants may allow identification of downstream genes.

At the current density of robotic printing, it is feasible to scale up the fabrication process to produce arrays containing 20,000 cDNA targets. At this density, a single array would be sufficient to provide gene-specific targets encompassing nearly the entire repertoire of expressed genes in the *Arabidopsis* genome (*2*). The availability of 20,274 ESTs from *Arabidopsis* (*1*, *9*) would provide a rich source of templates for such studies.

The estimated 100,000 genes in the human genome (*10*) exceeds the number of *Arabidopsis* genes by a factor of 5 (*2*). This modest increase in complexity suggests that similar cDNA microarrays, prepared from the rapidly growing repertoire of human ESTs (*1*), could be used to determine the expression patterns of tens of thousands of human genes in diverse cell types. Coupling an amplification strategy to the reverse transcription reaction (*11*) could make it feasible to monitor expression even in minute tissue samples. A wide variety of acute and chronic physiological and pathological conditions might lead to characteristic changes in the patterns of gene expression in peripheral blood cells or other easily sampled tissues. In concert with cDNA microarrays for monitoring complex expression patterns, these tissues might therefore serve as sensitive in vivo sensors for clinical diagnosis. Microarrays of cDNAs could thus provide a useful link between human gene sequences and clinical medicine.

**Table 1.** Sequences contained on the cDNA microarray. Shown is the position, the known or putative function, and the accession number of each cDNA in the microarray (Fig. 1). All but three of the ESTs used in this study matched a sequence in the database. NADH, reduced form of nicotinamide adenine dinucleotide; ATPase, adenosine triphosphatase; GTP, guanosine triphosphate.

| Position | cDNA | Function | Accession number |
|---|---|---|---|
| a1, 2 | AChR | Human AChR | * |
| a3, 4 | EST3 | Actin | H36236 |
| a5, 6 | EST6 | NADH dehydrogenase | Z27010 |
| a7, 8 | *AAC1* | Actin 1 | M20016 |
| a9, 10 | EST12 | Unknown | U36594† |
| a11, 12 | EST13 | Actin | T45783 |
| b1, 2 | *CAB1* | Chlorophyll a/b binding | M85150 |
| b3, 4 | EST17 | Phosphoglycerate kinase | T44490 |
| b5, 6 | *GA4* | Gibberellic acid biosynthesis | L37126 |
| b7, 8 | EST19 | Unknown | U36595† |
| b9, 10 | *GBF-1* | G-box binding factor 1 | X63894 |
| b11, 12 | EST23 | Elongation factor | X52256 |
| c1, 2 | EST29 | Aldolase | T04477 |
| c3, 4 | *GBF-2* | G-box binding factor 2 | X63895 |
| c5, 6 | EST34 | Chloroplast protease | R87034 |
| c7, 8 | EST35 | Unknown | T14152 |
| c9, 10 | EST41 | Catalase | T22720 |
| c11, 12 | rGR | Rat glucocorticoid receptor | M14053 |
| d1, 2 | EST42 | Unknown | U36596† |
| d3, 4 | EST45 | ATPase | J04185 |
| d5, 6 | *HAT1* | Homeobox-leucine zipper 1 | U09332 |
| d7, 8 | EST46 | Light harvesting complex | T04063 |
| d9, 10 | EST49 | Unknown | T76267 |
| d11, 12 | *HAT2* | Homeobox-leucine zipper 2 | U09335 |
| e1, 2 | *HAT4* | Homeobox-leucine zipper 4 | M90394 |
| e3, 4 | EST50 | Phosphoribulokinase | T04344 |
| e5, 6 | *HAT5* | Homeobox-leucine zipper 5 | M90416 |
| e7, 8 | EST51 | Unknown | Z33675 |
| e9, 10 | *HAT22* | Homeobox-leucine zipper 22 | U09336 |
| e11, 12 | EST52 | Oxygen evolving | T21749 |
| f1, 2 | EST59 | Unknown | Z34607 |
| f3, 4 | *KNAT1* | *Knotted*-like homeobox 1 | U14174 |
| f5, 6 | EST60 | RuBisCO small subunit | X14564 |
| f7, 8 | EST69 | Translation elongation factor | T42799 |
| f9, 10 | *PPH1* | Protein phosphatase 1 | U34803 |
| f11, 12 | EST70 | Unknown | T44621 |
| g1, 2 | EST75 | Chloroplast protease | T43698 |
| g3, 4 | EST78 | Unknown | R65481 |
| g5, 6 | *ROC1* | Cyclophilin | L14844 |
| g7, 8 | EST82 | GTP binding | X59152 |
| g9, 10 | EST83 | Unknown | Z33795 |
| g11, 12 | EST84 | Unknown | T45278 |
| h1, 2 | EST91 | Unknown | T13832 |
| h3, 4 | EST96 | Unknown | R64816 |
| h5, 6 | *SAR1* | Synaptobrevin | M90418 |
| h7, 8 | EST100 | Light harvesting complex | Z18205 |
| h9, 10 | EST103 | Light harvesting complex | X03909 |
| h11, 12 | *TRP4* | Yeast tryptophan biosynthesis | X04273 |

*Proprietary sequence of Stratagene (La Jolla, California). †No match in the database; novel EST.

**Table 2.** Gene expression monitoring by microarray and RNA blot analyses; tg, *HAT4*-transgenic. See Table 1 for additional gene information. Expression levels (w/w) were calibrated with the use of known amounts of human AChR mRNA. Values for the microarray were determined from microarray scans (Fig. 1); values for the RNA blot were determined from RNA blots (Fig. 2).

| Gene | Expression level (w/w) | |
|---|---|---|
| | Microarray | RNA blot |
| *CAB1* | 1:48 | 1:83 |
| *CAB1* (tg) | 1:120 | 1:150 |
| *HAT4* | 1:8300 | 1:6300 |
| *HAT4* (tg) | 1:150 | 1:210 |
| *ROC1* | 1:1200 | 1:1800 |
| *ROC1* (tg) | 1:260 | 1:1300 |

## REFERENCES AND NOTES


1. The current EST database (dbEST release 091495) from the National Center for Biotechnology Information (Bethesda, MD) contains a total of 322,225 entries, including 255,845 from the human genome and 21,044 from *Arabidopsis*. Access is available via the World Wide Web (http://www.ncbi.nlm.nih.gov).
2. E. M. Meyerowitz and R. E. Pruitt, *Science* 229, 1214 (1985); R. E. Pruitt and E. M. Meyerowitz, *J. Mol. Biol.* 187, 169 (1986); L. Hwang *et al.*, *Plant J.* 1, 367 (1991); P. Janis *et al.*, *Plant Mol. Biol.* 24, 685 (1994); L. Le Guen *et al.*, *Mol. Gen. Genet.* 245, 392 (1994).
3. D. Shalon, thesis, Stanford University (1995); ____ and P. O. Brown, in preparation. Microarrays were fabricated on poly-L-lysine–coated microscope slides (Sigma) with a custom-built arraying machine fitted with one printing tip. The tip loaded 1 μl of PCR product (0.5 mg/ml) from 96-well microtiter plates and deposited ~0.005 μl per slide on 40 slides at a spacing of 500 μm. The printed slides were rehydrated for 2 hours in a humid chamber, snap-dried at 100°C for 1 min, rinsed in 0.1% SDS, and treated with 0.05% succinic anhydride prepared in buffer consisting of 50% 1-methyl-2-pyrrolidinone and 50% boric acid. The cDNA on the slides was denatured in distilled water for 2 min at 90°C immediately before use. Microarrays were scanned with a laser fluorescent scanner that contained a computer-controlled XY stage and a microscope objective. A mixed gas, multiline laser allowed sequential excitation of the two fluorophores. Emitted light was split according to wavelength and detected with two photomultiplier tubes. Signals were read into a PC with the use of a 12-bit analog-to-digital board. Additional details of microarray fabrication and use may be obtained by means of e-mail (pbrown@cmgm.stanford.edu).
4. F. M. Ausubel *et al.*, Eds., *Current Protocols in Molecular Biology* (Greene & Wiley Interscience, New York, 1994), pp. 4.3.1–4.3.4.
5. Polyadenylated (poly(A)$^+$) mRNA was prepared from total RNA with the use of Oligotex-dT resin (Qiagen). Reverse transcription (RT) reactions were carried out with a StrataScript RT-PCR kit (Stratagene) modified as follows: 50-μl reactions contained 0.1 μg/μl of *Arabidopsis* mRNA, 0.1 ng/μl of human AChR mRNA, 0.05 μg/μl of oligo(dT) (21-mer), 1× first strand buffer, 0.03 U/μl of ribonuclease block, 500 μM deoxyadenosine triphosphate (dATP), 500 μM deoxyguanosine triphosphate, 500 μM dTTP, 40 μM deoxycytosine triphosphate (dCTP), 40 μM fluorescein-12-dCTP (or tetramine-5-dCTP), and 0.03 U/μl of StrataScript reverse transcriptase. Reactions were incubated for 60 min at 37°C, precipitated with ethanol, and resuspended in 10 μl of TE (10 mM tris-HCl and 1 mM EDTA, pH 8.0). Samples were then heated for 3 min at 94°C and chilled on ice. The RNA was degraded by adding 0.25 μl of 10 N NaOH followed by a 10-min incubation at 37°C. The samples were neutralized by addition of 2.5 μl of 1 M tris-Cl (pH 6.0) and 25 μl of 10 N HCl and precipitated with ethanol. Pellets were washed with 70% ethanol, dried to completion in a speedvac, resuspended in 10 μl of $H_2O$, and reduced to 3.0 μl in a speedvac. Fluorescent nucleotide analogs were obtained from New England Nuclear (DuPont).
6. Hybridization reactions contained 1.0 μl of fluorescent cDNA synthesis product (5) and 1.0 μl of hybridization buffer [10× saline sodium citrate (SSC) and 0.2% SDS]. The 2.0-μl probe mixtures were aliquoted onto the microarray surface and covered with cover slips (12 mm round). Arrays were transferred to a hybridization chamber (3) and incubated for 18 hours at 65°C. Arrays were washed for 5 min at room temperature (25°C) in low-stringency wash buffer (1× SSC and 0.1% SDS), then for 10 min at room temperature in high-stringency wash buffer (0.1× SSC and 0.1% SDS). Arrays were scanned in 0.1× SSC with the use of a fluorescence laser-scanning device (3).
7. Samples of poly(A)$^+$ mRNA (4, 5) were spotted onto nylon membranes (hybran) and crosslinked with ultraviolet light with the use of a Stratalinker 1800 (Stratagene). Probes were prepared by random priming with the use of a Prime-It II kit (Stratagene) in the presence of [$^{32}$P]dATP. Hybridizations were carried out according to the instructions of the manufacturer. Quantitation was performed on a PhosphorImager (Molecular Dynamics).
8. M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 89, 3894 (1992); M. Schena, A. M. Lloyd, R. W. Davis, *Genes Dev.* 7, 367 (1993); M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 91, 8393 (1994).
9. H. Hofte *et al.*, *Plant J.* 4, 1051 (1993); T. Newman *et al.*, *Plant Physiol.* 106, 1241 (1994).
10. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* 88, 7474 (1991); E. D. Green and R. H. Waterston, *J. Am. Med. Assoc.* 266, 1966 (1991); C. Belaiche-Chantelot, *Cell* 70, 1059 (1992); D. R. Cox *et al.*, *Science* 265, 2031 (1994).
11. E. S. Kawasaki *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 85, 5698 (1988).
12. The laser fluorescent scanner was designed and fabricated in collaboration with S. Smith of Stanford University. Scanner and analysis software was developed by R. X. Xia. The succinic anhydride reaction was suggested by J. Mulligan and J. Van Ness of Darwin Molecular Corporation. Thanks to S. Theologis, C. Somerville, K. Yamamoto, and members of the laboratories of R.W.D. and P.O.B. for critical comments. Supported by the Howard Hughes Medical Institute and by grants from NIH [R21HG00450 (P.O.B.) and R37AG00198 (R.W.D.)] and from NSF (MCB9106011) (R.W.D.) and by an NSF graduate fellowship (D.S.). P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.


11 August 1995; accepted 22 September 1995

# Gene Therapy in Peripheral Blood Lymphocytes and Bone Marrow for ADA$^-$ Immunodeficient Patients

Claudio Bordignon,* Luigi D. Notarangelo, Nadia Nobili, Giuliana Ferrari, Giulia Casorati, Paola Panina, Evelina Mazzolari, Daniela Maggioni, Claudia Rossi, Paolo Servida, Alberto G. Ugazio, Fulvio Mavilio

Adenosine deaminase (ADA) deficiency results in severe combined immunodeficiency, the first genetic disorder treated by gene therapy. Two different retroviral vectors were used to transfer ex vivo the human ADA minigene into bone marrow cells and peripheral blood lymphocytes from two patients undergoing exogenous enzyme replacement therapy. After 2 years of treatment, long-term survival of T and B lymphocytes, marrow cells, and granulocytes expressing the transferred ADA gene was demonstrated and resulted in normalization of the immune repertoire and restoration of cellular and humoral immunity. After discontinuation of treatment, T lymphocytes, derived from transduced peripheral blood lymphocytes, were progressively replaced by marrow-derived T cells in both patients. These results indicate successful gene transfer into long-lasting progenitor cells, producing a functional multilineage progeny.

Severe combined immunodeficiency associated with inherited deficiency of ADA (*1*) is usually fatal unless affected children are kept in protective isolation or the immune system is reconstituted by bone marrow transplantation from a human leukocyte antigen (HLA)–identical sibling donor (*2*). This is the therapy of choice, although it is available only for a minority of patients. In recent years, other forms of therapy have been developed, including transplants from haploidentical donors (*3*, *4*), exogenous enzyme replacement (*5*), and somatic-cell gene therapy (*6*–*9*).

We previously reported a preclinical model in which ADA gene transfer and expression successfully restored immune functions in human ADA-deficient (ADA$^-$) peripheral blood lymphocytes (PBLs) in immunodeficient mice in vivo (*10*, *11*). On the basis of these preclinical results, the clinical application of gene therapy for the treatment of ADA$^-$ SCID (severe combined immunodeficiency disease) patients who previously failed exogenous enzyme replacement therapy was approved by our Institutional Ethical Committees and by the Italian National Committee for Bioethics (*12*). In addition to evaluating the safety and efficacy of the gene therapy procedure, the aim of the study was to define the relative role of PBLs and hematopoietic stem cells in the long-term reconstitution of immune functions after retroviral vector–mediated ADA gene transfer. For this purpose, two structurally identical vectors expressing the human ADA complementary DNA (cDNA), distinguishable by the presence of alternative restriction sites in a nonfunctional region of the viral long-terminal repeat (LTR), were used to transduce PBLs and bone marrow (BM) cells independently. This procedure allowed identification of the origin of

C. Bordignon, N. Nobili, G. Ferrari, D. Maggioni, C. Rossi, P. Servida, F. Mavilio, Telethon Gene Therapy Program for Genetic Diseases, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
L. D. Notarangelo, E. Mazzolari, A. G. Ugazio, Department of Pediatrics, University of Brescia Medical School, Brescia, Italy.
G. Casorati, Unità di Immunochimica, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
P. Panina, Roche Milano Ricerche, Milan, Italy.
*To whom correspondence should be addressed.

**PCT** WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G01N 33/543, 33/68

A1

(11) International Publication Number: **WO 95/35505**

(43) International Publication Date: 28 December 1995 (28.12.95)

(21) International Application Number: PCT/US95/07659

(22) International Filing Date: 16 June 1995 (16.06.95)

(30) Priority Data:

| | | |
|---|---|---|
| 08/261,388 | 17 June 1994 (17.06.94) | US |
| 08/477,809 | 7 June 1995 (07.06.95) | US |

(71) Applicant: THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY [US/US]; Stanford, CA 94305 (US).

(72) Inventors: SHALON, Tidhar, Dari; 364 Fletcher Drive, Atherton, CA 94027 (US). BROWN, Patrick, O.; 76 Peter Coutts Circle, Stanford, CA 94305 (US).

(74) Agent: DEHLINGER, Peter, J.; Dehlinger & Associates, P.O. Box 60850, Palo Alto, CA 94306-1546 (US).

(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

**Published**
*With international search report.*

Reference 4d
In USSN: 09/838,044
PB-0011-1 DIV

(54) Title: METHOD AND APPARATUS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES

(57) Abstract

A method and apparatus for forming microarrays of biological samples on a support are disclosed. The method involves dispensing a known volume of a reagent at each of a selected array position, by tapping a capillary dispenser on the support under conditions effective to draw a defined volume of liquid onto the support. The apparatus is designed to produce a microarray of such regions in an automated fashion.

*FOR THE PURPOSES OF INFORMATION ONLY*

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

| | | | | | |
|---|---|---|---|---|---|
| AT | Austria | GB | United Kingdom | MR | Mauritania |
| AU | Australia | GE | Georgia | MW | Malawi |
| BB | Barbados | GN | Guinea | NE | Niger |
| BE | Belgium | GR | Greece | NL | Netherlands |
| BF | Burkina Faso | HU | Hungary | NO | Norway |
| BG | Bulgaria | IE | Ireland | NZ | New Zealand |
| BJ | Benin | IT | Italy | PL | Poland |
| BR | Brazil | JP | Japan | PT | Portugal |
| BY | Belarus | KE | Kenya | RO | Romania |
| CA | Canada | KG | Kyrgystan | RU | Russian Federation |
| CF | Central African Republic | KP | Democratic People's Republic of Korea | SD | Sudan |
| CG | Congo | | | SE | Sweden |
| CH | Switzerland | KR | Republic of Korea | SI | Slovenia |
| CI | Côte d'Ivoire | KZ | Kazakhstan | SK | Slovakia |
| CM | Cameroon | LI | Liechtenstein | SN | Senegal |
| CN | China | LK | Sri Lanka | TD | Chad |
| CS | Czechoslovakia | LU | Luxembourg | TG | Togo |
| CZ | Czech Republic | LV | Latvia | TJ | Tajikistan |
| DE | Germany | MC | Monaco | TT | Trinidad and Tobago |
| DK | Denmark | MD | Republic of Moldova | UA | Ukraine |
| ES | Spain | MG | Madagascar | US | United States of America |
| FI | Finland | ML | Mali | UZ | Uzbekistan |
| FR | France | MN | Mongolia | VN | Viet Nam |
| GA | Gabon | | | | |

# METHOD AND APPARATUS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES

## Field of the Invention

This invention relates to a method and apparatus for fabricating microarrays of biological samples for large scale screening assays, such as arrays of DNA samples to be used in DNA hybridization assays for genetic research and diagnostic applications.

## References

Abouzied, *et al.*, *Journal of AOAC International* 77(2):495-500 (1994).

Bohlander, *et al.*, *Genomics* 13:1322-1324 (1992).

Drmanac, *et al.*, *Science* 260:1649-1652 (1993).

Fodor, *et al.*, *Science* 251:767-773 (1991).

Khrapko, *et al.*, *DNA Sequence* 1:375-388 (1991).

Kuriyama, *et al.*, AN ISFET BIOSENSOR, APPLIED BIOSENSORS (Donald Wise, Ed.), Butterworths, pp. 93-114 (1989).

Lehrach, *et al.*, HYBRIDIZATION FINGERPRINTING IN GENOME MAPPING AND SEQUENCING, GENOME ANALYSIS, VOL 1 (Davies and Tilgham, Eds.), Cold Spring Harbor Press, pp. 39-81 (1990).

Maniatis, *et al.*, MOLECULAR CLONING, A LABORATORY MANUAL, Cold Spring Harbor Press (1989).

Nelson, *et al.*, *Nature Genetics* 4:11-18 (1993).

Pirrung, *et al.*, U.S. Patent No. 5,143,854 (1992).

Riles, *et al.*, *Genetics* 134:81-150 (1993).

Schena, M. *et al.*, *Proc. Nat. Acad. Sci. USA* 89:3894-3898 (1992).

Southern, *et al.*, *Genomics* 13:1008-1017 (1992).

## Background of the Invention

A variety of methods are currently available for making arrays of biological macromolecules, such as arrays of nucleic acid molecules or proteins. One method for making ordered arrays of DNA on a porous membrane is a "dot blot" approach. In this method, a vacuum manifold transfers a plurality, *e.g.*, 96, aqueous samples of DNA from 3 millimeter diameter wells to a porous membrane. A common variant of this procedure is a "slot-blot" method in which the wells have highly-elongated oval shapes.

The DNA is immobilized on the porous membrane by baking the membrane or exposing it to UV radiation. This is a manual procedure practical for making one array at a time and usually limited to 96 samples per array. "Dot-blot" procedures are therefore inadequate for applications in which many thousand samples must be determined.

A more efficient technique employed for making ordered arrays of genomic fragments uses an array of pins dipped into the wells, *e.g.*, the 96 wells of a microtitre plate, for transferring an array of samples to a substrate, such as a porous membrane. One array includes pins that are designed to spot a membrane in a staggered fashion, for creating an array of 9216 spots in a 22 × 22 cm area (Lehrach, *et al.*, 1990). A limitation with this approach is that the volume of DNA spotted in each pixel of each array is highly variable.

In addition, the number of arrays that can be made with each dipping is usually quite small.

An alternate method of creating ordered arrays of nucleic acid sequences is described by Pirrung, *et al.* (1992), and also by Fodor, *et al.* (1991). The method involves synthesizing different nucleic acid sequences at different discrete regions of a support. This method employs elaborate synthetic schemes, and is generally limited to relatively short nucleic acid sample, *e.g.*, less than 20 bases. A related method has been described by Southern, *et al.* (1992).

Khrapko, *et al.* (1991) describes a method of making an oligonucleotide matrix by spotting DNA onto a thin layer of polyacrylamide. The spotting is done manually with a micropipette.

None of the methods or devices described in the prior art are designed for mass fabrication of microarrays characterized by (i) a large number of micro-sized assay regions separated by a distance of 50-200 microns or less, and (ii) a well-defined amount, typically in the picomole range, of analyte associated with each region of the array.

Furthermore, current technology is directed at performing such assays one at a time to a single array of DNA molecules. For example, the most common method for performing DNA hybridizations to arrays spotted onto porous membrane involves sealing the membrane in a plastic bag (Maniatas, *et al.*, 1989) or a rotating glass cylinder (Robbins Scientific) with the labeled hybridization probe inside the sealed chamber. For arrays made on non-porous surfaces, such as a microscope slide, each array is incubated with the labeled hybridization probe sealed under a coverslip. These techniques require a separate sealed chamber for

each array which makes the screening and handling of many such arrays inconvenient and time intensive.

Abouzied, *et al.* (1994) describes a method of printing horizontal lines of antibodies on a nitrocellulose membrane and separating regions of the membrane with vertical stripes of a hydrophobic material. Each vertical stripe is then reacted with a different antigen and the reaction between the immobilized antibody and an antigen is detected using a standard ELISA colorimetric technique. Abouzied's technique makes it possible to screen many one-dimensional arrays simultaneously on a single sheet of nitrocellulose. Abouzied makes the nitrocellulose somewhat hydrophobic using a line drawn with PAP Pen (Research Products International). However Abouzied does not describe a technology that is capable of completely sealing the pores of the nitrocellulose. The pores of the nitrocellulose are still physically open and so the assay reagents can leak through the hydrophobic barrier during extended high temperature incubations or in the presence of detergents which makes the Abouzied technique unacceptable for DNA hybridization assays.

Porous membranes with printed patterns of hydrophilic/hydrophobic regions exist for applications such as ordered arrays of bacteria colonies. QA Life Sciences (San Diego CA) makes such a membrane with a grid pattern printed on it. However, this membrane has the same disadvantage as the Abouzied technique since reagents can still flow between the gridded arrays making them unusable for separate DNA hybridization assays.

Pall Corporation make a 96-well plate with a porous filter heat sealed to the bottom of the plate. These plates are capable of containing different

reagents in each well without cross-contamination. However, each well is intended to hold only one target element whereas the invention described here makes a microarray of many biomolecules in each subdivided region of the solid support. Furthermore, the 96 well plates are at least 1 cm thick and prevent the use of the device for many colorimetric, fluorescent and radioactive detection formats which require that the membrane lie flat against the detection surface. The invention described here requires no further processing after the assay step since the barriers elements are shallow and do not interfere with the detection step thereby greatly increasing convenience.

Hyseq Corporation has described a method of making an "array of arrays" on a non-porous solid support for use with their sequencing by hybridization technique. The method described by Hyseq involves modifying the chemistry of the solid support material to form a hydrophobic grid pattern where each subdivided region contains a microarray of biomolecules. Hyseq's flat hydrophobic pattern does not make use of physical blocking as an additional means of preventing cross contamination.

## Summary of the Invention

The invention includes, in one aspect, a method of forming a microarray of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent. The method involves first loading a solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel (i) formed by spaced-apart, coextensive elongate members, (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous

solution in the channel forms a meniscus. The channel is preferably formed by a pair of spaced-apart tapered elements.

The tip of the dispensing device is tapped against a solid support at a defined position on the support surface with an impulse effective to break the meniscus in the capillary channel deposit a selected volume of solution on the surface, preferably a selected volume in the range 0.01 to 100 nl. The two steps are repeated until the desired array is formed.

The method may be practiced in forming a plurality of such arrays, where the solution-depositing step is are applied to a selected position on each of a plurality of solid supports at each repeat cycle.

The dispensing device may be loaded with a new solution, by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

Also included in the invention is an automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected, analyte-specific reagent. The apparatus has a holder for holding, at known positions, a plurality of planar supports, and a reagent dispensing device of the type described above.

The apparatus further includes positioning structure for positioning the dispensing device at a selected array position with respect to a support in said holder, and dispensing structure for moving the dispensing device into tapping engagement against a support with a selected impulse effective to deposit a

selected volume on the support, e.g., a selected volume in the volume range 0.01 to 100 nl.

The positioning and dispensing structures are controlled by a control unit in the apparatus. The unit operates to (i) place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and (iii) dispense the reagent at a defined array position on each of the supports on said holder. The unit may further operate, at the end of a dispensing cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) remove the wash fluid prior to loading the dispensing device with a fresh selected reagent.

The dispensing device in the apparatus may be one of a plurality of such devices which are carried on the arm for dispensing different analyte assay reagents at selected spaced array positions.

In another aspect, the invention includes a substrate with a surface having a microarray of at least $10^3$ distinct polynucleotide or polypeptide biopolymers in a surface area of less than about 1 $cm^2$. Each distinct biopolymer (i) is disposed at a separate, defined position in said array, (ii) has a length of at least 50 subunits, and (iii) is present in a defined amount between about 0.1 femtomoles and 100 nanomoles.

In one embodiment, the surface is glass slide surface coated with a polycationic polymer, such as polylysine, and the biopolymers are polynucleotides. In another embodiment, the substrate has a water-impermeable backing, a water-permeable film formed on

the backing, and a grid formed on the film. The grid is composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film, and partitions the film into a plurality of water-impervious cells. A biopolymer array is formed within each well.

More generally, there is provided a substrate for use in detecting binding of labeled polynucleotides to one or more of a plurality different-sequence, immobilized polynucleotides. The substrate includes, in one aspect, a glass support, a coating of a polycationic polymer, such as polylysine, on said surface of the support, and an array of distinct polynucleotides electrostatically bound non-covalently to said coating, where each distinct biopolymer is disposed at a separate, defined position in a surface array of polynucleotides.

In another aspect, the substrate includes a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film, where the grid is composed of intersecting water-impervious grid elements extending from the backing to positions raised above the surface of the film, forming a plurality of cells. A biopolymer array is formed within each cell.

Also forming part of the invention is a method of detecting differential expression of each of a plurality of genes in a first cell type, with respect to expression of the same genes in a second cell type. In practicing the method, there is first produced fluorescent-labeled cDNA's from mRNA's isolated from the two cells types, where the cDNA'S from the first and second cells are labeled with first and second different fluorescent reporters.

A mixture of the labeled cDNA's from the two cell types is added to an array of polynucleotides

representing a plurality of known genes derived from the two cell types, under conditions that result in hybridization of the cDNA's to complementary-sequence polynucleotides in the array. The array is then examined by fluorescence under fluorescence excitation conditions in which (i) polynucleotides in the array that are hybridized predominantly to cDNA's derived from one of the first and second cell types give a distinct first or second fluorescence emission color, respectively, and (ii) polynucleotides in the array that are hybridized to substantially equal numbers of cDNA's derived from the first and second cell types give a distinct combined fluorescence emission color, respectively. The relative expression of known genes in the two cell types can then be determined by the observed fluorescence emission color of each spot.

These and other objects and features of the invention will become more fully apparent when the following detailed description of the invention is read in conjunction with the accompanying figures.

## Brief Description of the Drawings

Fig. 1 is a side view of a reagent-dispensing device having a open-capillary dispensing head constructed for use in one embodiment of the invention;

Figs. 2A-2C illustrate steps in the delivery of a fixed-volume bead on a hydrophobic surface employing the dispensing head from Fig. 1, in accordance with one embodiment of the method of the invention;

Fig. 3 shows a portion of a two-dimensional array of analyte-assay regions constructed according to the method of the invention;

Fig. 4 is a planar view showing components of an automated apparatus for forming arrays in accordance with the invention.

Fig. 5 shows a fluorescent image of an actual 20 × 20 array of 400 fluorescently-labeled DNA samples immobilized on a poly-l-lysine coated slide, where the total area covered by the 400 element array is 16 square millimeters;

Fig. 6 is a fluorescent image of a 1.8 cm × 1.8 cm microarray containing lambda clones with yeast inserts, the fluorescent signal arising from the hybridization to the array with approximately half the yeast genome labeled with a green fluorophore and the other half with a red fluorophore;

Fig. 7 shows the translation of the hybridization image of Fig. 6 into a karyotype of the yeast genome, where the elements of Fig.-6 microarray contain yeast DNA sequences that have been previously physically mapped in the yeast genome;

Fig. 8 show a fluorescent image of a 0.5 cm × 0.5 cm microarray of 24 cDNA clones, where the microarray was hybridized simultaneously with total cDNA from wild type *Arabidopsis* plant labeled with a green fluorophore and total cDNA from a transgenic *Arabidopsis* plant labeled with a red fluorophore, and the arrow points to the cDNA clone representing the gene introduced into the transgenic *Arabidopsis* plant;

Fig. 9 shows a plan view of substrate having an array of cells formed by barrier elements in the form of a grid;

Fig. 10 shows an enlarged plan view of one of the cells in the substrate in Fig. 9, showing an array of polynucleotide regions in the cell;

Fig. 11 is an enlarged sectional view of the substrate in Fig. 9, taken along a section line in that figure; and

Fig. 12 is a scanned image of a 3 cm × 3 cm nitrocellulose solid support containing four identical

arrays of M13 clones in each of four quadrants, where each quadrant was hybridized simultaneously to a different oligonucleotide using an open face hybridization method.

## Detailed Description of the Invention

### I. Definitions

Unless indicated otherwise, the terms defined below have the following meanings:

"Ligand" refers to one member of a ligand/anti-ligand binding pair. The ligand may be, for example, one of the nucleic acid strands in a complementary, hybridized nucleic acid duplex binding pair; an effector molecule in an effector/receptor binding pair; or an antigen in an antigen/antibody or antigen/antibody fragment binding pair.

"Antiligand" refers to the opposite member of a ligand/anti-ligand binding pair. The antiligand may be the other of the nucleic acid strands in a complementary, hybridized nucleic acid duplex binding pair; the receptor molecule in an effector/receptor binding pair; or an antibody or antibody fragment molecule in antigen/antibody or antigen/antibody fragment binding pair, respectively.

"Analyte" or "analyte molecule" refers to a molecule, typically a macromolecule, such as a polynucleotide or polypeptide, whose presence, amount, and/or identity are to be determined. The analyte is one member of a ligand/anti-ligand pair.

"Analyte-specific assay reagent" refers to a molecule effective to bind specifically to an analyte molecule. The reagent is the opposite member of a ligand/anti-ligand binding pair.

An "array of regions on a solid support" is a linear or two-dimensional array of preferably discrete

regions, each having a finite area, formed on the surface of a solid support.

A "microarray" is an array of regions having a density of discrete regions of at least about 100/cm$^2$, and preferably at least about 1000/cm$^2$. The regions in a microarray have typical dimensions, e.g., diameters, in the range of between about 10-250 μm, and are separated from other regions in the array by about the same distance.

A support surface is "hydrophobic" if a aqueous-medium droplet applied to the surface does not spread out substantially beyond the area size of the applied droplet. That is, the surface acts to prevent spreading of the droplet applied to the surface by hydrophobic interaction with the droplet.

A "meniscus" means a concave or convex surface that forms on the bottom of a liquid in a channel as a result of the surface tension of the liquid.

"Distinct biopolymers", as applied to the biopolymers forming a microarray, means an array member which is distinct from other array members on the basis of a different biopolymer sequence, and/or different concentrations of the same or distinct biopolymers, and/or different mixtures of distinct or different-concentration biopolymers. Thus an array of "distinct polynucleotides" means an array containing, as its members, (i) distinct polynucleotides, which may have a defined amount in each member, (ii) different, graded concentrations of given-sequence polynucleotides, and/or (iii) different-composition mixtures of two or more distinct polynucleotides.

"Cell type" means a cell from a given source, e.g., a tissue, or organ, or a cell in a given state of

differentiation, or a cell associated with a given pathology or genetic makeup.

II. Method of Microarray Formation

This section describes a method of forming a microarray of analyte-assay regions on a solid support or substrate, where each region in the array has a known amount of a selected, analyte-specific reagent.

Fig. 1 illustrates, in a partially schematic view, a reagent-dispensing device 10 useful in practicing the method. The device generally includes a reagent dispenser 12 having an elongate open capillary channel 14 adapted to hold a quantity of the reagent solution, such as indicated at 16, as will be described below. The capillary channel is formed by a pair of spaced-apart, coextensive, elongate members 12a, 12b which are tapered toward one another and converge at a tip or tip region 18 at the lower end of the channel. More generally, the open channel is formed by at least two elongate, spaced-apart members adapted to hold a quantity of reagent solutions and having a tip region at which aqueous solution in the channel forms a meniscus, such as the concave meniscus illustrated at 20 in Fig. 2A. The advantages of the open channel construction of the dispenser are discussed below.

With continued reference to Fig. 1, the dispenser device also includes structure for moving the dispenser rapidly toward and away from a support surface, for effecting deposition of a known amount of solution in the dispenser on a support, as will be described below with reference to Figs. 2A-2C. In the embodiment shown, this structure includes a solenoid 22 which is activatable to draw a solenoid piston 24 rapidly downwardly, then release the piston, *e.g.*, under spring bias, to a normal, raised position, as shown. The

dispenser is carried on the piston by a connecting member 26, as shown. The just-described moving structure is also referred to herein as dispensing means for moving the dispenser into engagement with a solid support, for dispensing a known volume of fluid on the support.

The dispensing device just described is carried on an arm 28 that may be moved either linearly or in an x-y plane to position the dispenser at a selected deposition position, as will be described.

Figs. 2A-2C illustrate the method of depositing a known amount of reagent solution in the just-described dispenser on the surface of a solid support, such as the support indicated at 30. The support is a polymer, glass, or other solid-material support having a surface indicated at 31.

In one general embodiment, the surface is a relatively hydrophilic, *i.e.*, wettable surface, such as a surface having native, bound or covalently attached charged groups. On such surface described below is a glass surface having an absorbed layer of a polycationic polymer, such as poly-l-lysine.

In another embodiment, the surface has or is formed to have a relatively hydrophobic character, *i.e.*, one that causes aqueous medium deposited on the surface to bead. A variety of known hydrophobic polymers, such as polystyrene, polypropylene, or polyethylene have desired hydrophobic properties, as do glass and a variety of lubricant or other hydrophobic films that may be applied to the support surface.

Initially, the dispenser is loaded with a selected analyte-specific reagent solution, such as by dipping the dispenser tip, after washing, into a solution of the reagent, and allowing filling by capillary flow into the dispenser channel. The dispenser is now moved

to a selected position with respect to a support surface, placing the dispenser tip directly above the support-surface position at which the reagent is to be deposited. This movement takes place with the dispenser tip in its raised position, as seen in Fig. 2A, where the tip is typically at least several 1-5 mm above the surface of the substrate.

With the dispenser so positioned, solenoid 22 is now activated to cause the dispenser tip to move rapidly toward and away from the substrate surface, making momentary contact with the surface, in effect, tapping the tip of the dispenser against the support surface. The tapping movement of the tip against the surface acts to break the liquid meniscus in the tip channel, bringing the liquid in the tip into contact with the support surface. This, in turn, produces a flowing of the liquid into the capillary space between the tip and the surface, acting to draw liquid out of the dispenser channel, as seen in Fig. 2B.

Fig. 2C shows flow of fluid from the tip onto the support surface, which in this case is a hydrophobic surface. The figure illustrates that liquid continues to flow from the dispenser onto the support surface until it forms a liquid bead 32. At a given bead size, *i.e.*, volume, the tendency of liquid to flow onto the surface will be balanced by the hydrophobic surface interaction of the bead with the support surface, which acts to limit the total bead area on the surface, and by the surface tension of the droplet, which tends toward a given bead curvature. At this point, a given bead volume will have formed, and continued contact of the dispenser tip with the bead, as the dispenser tip is being withdrawn, will have little or no effect on bead volume.

For liquid-dispensing on a more hydrophilic surface, the liquid will have less of a tendency to bead, and the dispensed volume will be more sensitive to the total dwell time of the dispenser tip in the immediate vicinity of the support surface, *e.g.*, the positions illustrated in Figs. 2B and 2C.

The desired deposition volume, *i.e.*, bead volume, formed by this method is preferably in the range 2 pl (picoliters) to 2 nl (nanoliters), although volumes as high as 100 nl or more may be dispensed. It will be appreciated that the selected dispensed volume will depend on (i) the "footprint" of the dispenser tip, *i.e.*, the size of the area spanned by the tip, (ii) the hydrophobicity of the support surface, and (iii) the time of contact with and rate of withdrawal of the tip from the support surface. In addition, bead size may be reduced by increasing the viscosity of the medium, effectively reducing the flow time of liquid from the dispenser onto the support surface. The drop size may be further constrained by depositing the drop in a hydrophilic region surrounded by a hydrophobic grid pattern on the support surface.

In a typical embodiment, the dispenser tip is tapped rapidly against the support surface, with a total residence time in contact with the support of less than about 1 msec, and a rate of upward travel from the surface of about 10 cm/sec.

Assuming that the bead that forms on contact with the surface is a hemispherical bead, with a diameter approximately equal to the width of the dispenser tip, as shown in Fig. 2C, the volume of the bead formed in relation to dispenser tip width (d) is given in Table 1 below. As seen, the volume of the bead ranges between 2 pl to 2 nl as the width size is increased from about 20 to 200 μm.

Table 1

| d | Volume (nl) |
|---|---|
| 20 μm | $2 \times 10^{-3}$ |
| 50 μm | $3.1 \times 10^{-2}$ |
| 100 μm | $2.5 \times 10^{-1}$ |
| 200 μm | 2 |

At a given tip size, bead volume can be reduced in a controlled fashion by increasing surface hydrophobicity, reducing time of contact of the tip with the surface, increasing rate of movement of the tip away from the surface, and/or increasing the viscosity of the medium. Once these parameters are fixed, a selected deposition volume in the desired pl to nl range can be achieved in a repeatable fashion.

After depositing a bead at one selected location on a support, the tip is typically moved to a corresponding position on a second support, a droplet is deposited at that position, and this process is repeated until a liquid droplet of the reagent has been deposited at a selected position on each of a plurality of supports.

The tip is then washed to remove the reagent liquid, filled with another reagent liquid and this reagent is now deposited at each another array position on each of the supports. In one embodiment, the tip is washed and refilled by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

From the foregoing, it will be appreciated that the tweezers-like, open-capillary dispenser tip

provides the advantages that (i) the open channel of the tip facilitates rapid, efficient washing and drying before reloading the tip with a new reagent, (ii) passive capillary action can load the sample directly from a standard microwell plate while retaining sufficient sample in the open capillary reservoir for the printing of numerous arrays, (iii) open capillaries are less prone to clogging than closed capillaries, and (iv) open capillaries do not require a perfectly faced bottom surface for fluid delivery.

A portion of a microarray 36 formed on the surface 38 of a solid support 40 in accordance with the method just described is shown in Fig. 3. The array is formed of a plurality of analyte-specific reagent regions, such as regions 42, where each region may include a different analyte-specific reagent. As indicated above, the diameter of each region is preferably between about 20-200 μm. The spacing between each region and its closest (non-diagonal) neighbor, measured from center-to-center (indicated at 44), is preferably in the range of about 20-400 μm. Thus, for example, an array having a center-to-center spacing of about 250 μm contains about 40 regions/cm or 1,600 regions/$cm^2$. After formation of the array, the support is treated to evaporate the liquid of the droplet forming each region, to leave a desired array of dried, relatively flat regions. This drying may be done by heating or under vacuum.

In some cases, it is desired to first rehydrate the droplets containing the analyte reagents to allow for more time for adsorption to the solid support. It is also possible to spot out the analyte reagents in a humid environment so that droplets do not dry until the arraying operation is complete.

III. Automated Apparatus for Forming Arrays

In another aspect, the invention includes an automated apparatus for forming an array of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent.

The apparatus is shown in planar, and partially schematic view in Fig. 4. A dispenser device 72 in the apparatus has the basic construction described above with respect to Fig. 1, and includes a dispenser 74 having an open-capillary channel terminating at a tip, substantially as shown in Figs. 1 and 2A-2C.

The dispenser is mounted in the device for movement toward and away from a dispensing position at which the tip of the dispenser taps a support surface, to dispense a selected volume of reagent solution, as described above. This movement is effected by a solenoid 76 as described above. Solenoid 76 is under the control of a control unit 77 whose operation will be described below. The solenoid is also referred to herein as dispensing means for moving the device into tapping engagement with a support, when the device is positioned at a defined array position with respect to that support.

The dispenser device is carried on an arm 74 which is threadedly mounted on a worm screw 80 driven (rotated) in a desired direction by a stepper motor 82 also under the control of unit 77. At its left end in the figure screw 80 is carried in a sleeve 84 for rotation about the screw axis. At its other end, the screw is mounted to the drive shaft of the stepper motor, which in turn is carried on a sleeve 86. The dispenser device, worm screw, the two sleeves mounting the worm screw, and the stepper motor used in moving the device in the "x" (horizontal) direction in the

figure form what is referred to here collectively as a displacement assembly 86.

The displacement assembly is constructed to produce precise, micro-range movement in the direction of the screw, *i.e.*, along an x axis in the figure. In one mode, the assembly functions to move the dispenser in x-axis increments having a selected distance in the range 5-25 μm. In another mode, the dispenser unit may be moved in precise x-axis increments of several microns or more, for positioning the dispenser at associated positions on adjacent supports, as will be described below.

The displacement assembly, in turn, is mounted for movement in the "y" (vertical) axis of the figure, for positioning the dispenser at a selected y axis position. The structure mounting the assembly includes a fixed rod 88 mounted rigidly between a pair of frame bars 90, 92, and a worm screw 94 mounted for rotation between a pair of frame bars 96, 98. The worm screw is driven (rotated) by a stepper motor 100 which operates under the control of unit 77. The motor is mounted on bar 96, as shown.

The structure just described, including worm screw 94 and motor 100, is constructed to produce precise, micro-range movement in the direction of the screw, *i.e.*, along an y axis in the figure. As above, the structure functions in one mode to move the dispenser in y-axis increments having a selected distance in the range 5-250 μm, and in a second mode, to move the dispenser in precise y-axis increments of several microns (μm) or more, for positioning the dispenser at associated positions on adjacent supports.

The displacement assembly and structure for moving this assembly in the y axis are referred to herein collectively as positioning means for positioning the

dispensing device at a selected array position with respect to a support.

A holder 102 in the apparatus functions to hold a plurality of supports, such as supports 104 on which the microarrays of regent regions are to be formed by the apparatus. The holder provides a number of recessed slots, such as slot 106, which receive the supports, and position them at precise selected positions with respect to the frame bars on which the dispenser moving means is mounted.

As noted above, the control unit in the device functions to actuate the two stepper motors and dispenser solenoid in a sequence designed for automated operation of the apparatus in forming a selected microarray of reagent regions on each of a plurality of supports.

The control unit is constructed, according to conventional microprocessor control principles, to provide appropriate signals to each of the solenoid and each of the stepper motors, in a given timed sequence and for appropriate signalling time. The construction of the unit, and the settings that are selected by the user to achieve a desired array pattern, will be understood from the following description of a typical apparatus operation.

Initially, one or more supports are placed in one or more slots in the holder. The dispenser is then moved to a position directly above a well (not shown) containing a solution of the first reagent to be dispensed on the support(s). The dispenser solenoid is actuated now to lower the dispenser tip into this well, causing the capillary channel in the dispenser to fill. Motors 82, 100 are now actuated to position the dispenser at a selected array position at the first of the supports. Solenoid actuation of the dispenser is

then effective to dispense a selected-volume droplet of that reagent at this location. As noted above, this operation is effective to dispense a selected volume preferably between 2 pl and 2 nl of the reagent solution.

The dispenser is now moved to the corresponding position at an adjacent support and a similar volume of the solution is dispensed at this position. The process is repeated until the reagent has been dispensed at this preselected corresponding position on each of the supports.

Where it is desired to dispense a single reagent at more than two array positions on a support, the dispenser may be moved to different array positions at each support, before moving the dispenser to a new support, or solution can be dispensed at individual positions on each support, at one selected position, then the cycle repeated for each new array position.

To dispense the next reagent, the dispenser is positioned over a wash solution (not shown), and the dispenser tip is dipped in and out of this solution until the reagent solution has been substantially washed from the tip. Solution can be removed from the tip, after each dipping, by vacuum, compressed air spray, sponge, or the like.

The dispenser tip is now dipped in a second reagent well, and the filled tip is moved to a second selected array position in the first support. The process of dispensing reagent at each of the corresponding second-array positions is then carried as above. This process is repeated until an entire microarray of reagent solutions on each of the supports has been formed.

## IV. Microarray Substrate

This section describes embodiments of a substrate having a microarray of biological polymers carried on the substrate surface. Subsection A describes a multi-cell substrate, each cell of which contains a microarray, and preferably an identical microarray, of distinct biopolymers, such as distinct polynucleotides, formed on a porous surface. Subsection B describes a microarray of distinct polynucleotides bound on a glass slide coated with a polycationic polymer.

A. Multi-Cell Substrate

Fig. 9 illustrates, in plan view, a substrate 110 constructed according to the invention. The substrate has an 8 x 12 rectangular array 112 of cells, such as cells 114, 116, formed on the substrate surface. With reference to Fig. 10, each cell, such as cell 114, in turn supports a microarray 118 of distinct biopolymers, such as polypeptides or polynucleotides at known, addressable regions of the microarray. Two such regions forming the microarray are indicated at 120, and correspond to regions, such as regions 42, forming the microarray of distinct biopolymers shown in Fig. 3.

The 96-cell array shown in Fig. 9 has typically array dimensions between about 12 and 244 mm in width and 8 and 400 mm in length, with the cells in the array having width and length dimension of 1/12 and 1/8 the array width and length dimensions, respectively, *i.e.*, between about 1 and 20 in width and 1 and 50 mm in length.

The construction of substrate is shown cross-sectionally in Fig. 11, which is an enlarged sectional view taken along view line 124 in Fig. 9. The substrate includes a water-impermeable backing 126, such as a glass slide or rigid polymer sheet. Formed on the surface of the backing is a water-permeable film

128. The film is formed of a porous membrane material, such as nitrocellulose membrane, or a porous web material, such as a nylon, polypropylene, or PVDF porous polymer material. The thickness of the film is preferably between about 10 and 1000 μm. The film may be applied to the backing by spraying or coating uncured material on the backing, or by applying a preformed membrane to the backing. The backing and film may be obtained as a preformed unit from commercial source, *e.g.*, a plastic-backed nitrocellulose film available from Schleicher and Schuell Corporation.

With continued reference to Fig. 11, the film-covered surface in the substrate is partitioned into a desired array of cells by water-impermeable grid lines, such as lines 130, 132, which have infiltrated the film down to the level of the backing, and extend above the surface of the film as shown, typically a distance of 100 to 2000 μm above the film surface.

The grid lines are formed on the substrate by laying down an uncured or otherwise flowable resin or elastomer solution in an array grid, allowing the material to infiltrate the porous film down to the backing, then curing or otherwise hardening the grid lines to form the cell-array substrate.

One preferred material for the grid is a flowable silicone available from Loctite Corporation. The barrier material can be extruded through a narrow syringe (*e.g.*, 22 gauge) using air pressure or mechanical pressure. The syringe is moved relative to the solid support to print the barrier elements as a grid pattern. The extruded bead of silicone wicks into the pores of the solid support and cures to form a shallow waterproof barrier separating the regions of the solid support.

In alternative embodiments, the barrier element can be a wax-based material or a thermoset material such as epoxy. The barrier material can also be a UV-curing polymer which is exposed to UV light after being printed onto the solid support. The barrier material may also be applied to the solid support using printing techniques such as silk-screen printing. The barrier material may also be a heat-seal stamping of the porous solid support which seals its pores and forms a water-impervious barrier element. The barrier material may also be a shallow grid which is laminated or otherwise adhered to the solid support.

In addition to plastic-backed nitrocellulose, the solid support can be virtually any porous membrane with or without a non-porous backing. Such membranes are readily available from numerous vendors and are made from nylon, PVDF, polysulfone and the like. In an alternative embodiment, the barrier element may also be used to adhere the porous membrane to a non-porous backing in addition to functioning as a barrier to prevent cross contamination of the assay reagents.

In an alternative embodiment, the solid support can be of a non-porous material. The barrier can be printed either before or after the microarray of biomolecules is printed on the solid support.

As can be appreciated, the cells formed by the grid lines and the underlying backing are water-impermeable, having side barriers projecting above the porous film in the cells. Thus, defined-volume samples can be placed in each well without risk of cross-contamination with sample material in adjacent cells. In Fig. 11, defined volumes samples, such as sample 134, are shown in the cells.

As noted above, each well contains a microarray of distinct biopolymers. In one general embodiment, the

microarrays in the well are identical arrays of distinct biopolymers, *e.g.*, different sequence polynucleotides. Such arrays can be formed in accordance with the methods described in Section II, by depositing a first selected polynucleotide at the same selected microarray position in each of the cells, then depositing a second polynucleotide at a different microarray position in each well, and so on until a complete, identical microarray is formed in each cell.

In a preferred embodiment, each microarray contains about $10^3$ distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 $cm^2$. Also in a preferred embodiment, the biopolymers in each microarray region are present in a defined amount between about 0.1 femtomoles and 100 nanomoles. The ability to form high-density arrays of biopolymers, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method described in Section II.

Also in a preferred embodiments, the biopolymers are polynucleotides having lengths of at least about 50 bp, *i.e.*, substantially longer than oligonucleotides which can be formed in high-density arrays by schemes involving parallel, step-wise polymer synthesis on the array surface.

In the case of a polynucleotide array, in an assay procedure, a small volume of the labeled DNA probe mixture in a standard hybridization solution is loaded onto each cell. The solution will spread to cover the entire microarray and stop at the barrier elements. The solid support is then incubated in a humid chamber at the appropriate temperature as required by the assay.

Each assay may be conducted in an "open-face" format where no further sealing step is required, since the hybridization solution will be kept properly hydrated by the water vapor in the humid chamber. At the conclusion of the incubation step, the entire solid support containing the numerous microarrays is rinsed quickly enough to dilute the assay reagents so that no significant cross contamination occurs. The entire solid support is then reacted with detection reagents if needed and analyzed using standard colorimetric, radioactive or fluorescent detection means. All processing and detection steps are performed simultaneously to all of the microarrays on the solid support ensuring uniform assay conditions for all of the microarrays on the solid support.

B. Glass-Slide Polynucleotide Array

Fig. 5 shows a substrate 136 formed according to another aspect of the invention, and intended for use in detecting binding of labeled polynucleotides to one or more of a plurality distinct polynucleotides. The substrate includes a glass substrate 138 having formed on its surface, a coating of a polycationic polymer, preferably a cationic polypeptide, such as polylysine or polyarginine. Formed on the polycationic coating is a microarray 140 of distinct polynucleotides, each localized at known selected array regions, such as regions 142.

The slide is coated by placing a uniform-thickness film of a polycationic polymer, e.g., poly-l-lysine, on the surface of a slide and drying the film to form a dried coating. The amount of polycationic polymer added is sufficient to form at least a monolayer of polymers on the glass surface. The polymer film is bound to surface via electrostatic binding between

negative silyl-OH groups on the surface and charged amine groups in the polymers. Poly-l-lysine coated glass slides may be obtained commercially, e.g., from Sigma Chemical Co. (St. Louis, MO).

To form the microarray, defined volumes of distinct polynucleotides are deposited on the polymer-coated slide, as described in Section II. According to an important feature of the substrate, the deposited polynucleotides remain bound to the coated slide surface non-covalently when an aqueous DNA sample is applied to the substrate under conditions which allow hybridization of reporter-labeled polynucleotides in the sample to complementary-sequence (single-stranded) polynucleotides in the substrate array. The method is illustrated in Examples 1 and 2.

To illustrate this feature, a substrate of the type just described, but having an array of same-sequence polynucleotides, was mixed with fluorescent-labeled complementary DNA under hybridization conditions. After washing to remove non-hybridized material, the substrate was examined by low-power fluorescence microscopy. The array can be visualized by the relatively uniform labeling pattern of the array regions.

In a preferred embodiment, each microarray contains at least $10^3$ distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 $cm^2$. In the embodiment shown in Fig. 5, the microarray contains 400 regions in an area of about 16 $mm^2$, or $2.5 \times 10^3$ regions/$cm^2$. Also in a preferred embodiment, the polynucleotides in the each microarray region are present in a defined amount between about 0.1 femtomoles and 100 nanomoles in the case of polynucleotides. As above, the ability to form high-

density arrays of this type, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method described in Section II.

Also in a preferred embodiments, the polynucleotides have lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by various in situ synthesis schemes.

V. Utility

Microarrays of immobilized nucleic acid sequences prepared in accordance with the invention can be used for large scale hybridization assays in numerous genetic applications, including genetic and physical mapping of genomes, monitoring of gene expression, DNA sequencing, genetic diagnosis, genotyping of organisms, and distribution of DNA reagents to researchers.

For gene mapping, a gene or a cloned DNA fragment is hybridized to an ordered array of DNA fragments, and the identity of the DNA elements applied to the array is unambiguously established by the pixel or pattern of pixels of the array that are detected. One application of such arrays for creating a genetic map is described by Nelson, *et al.* (1993). In constructing physical maps of the genome, arrays of immobilized cloned DNA fragments are hybridized with other cloned DNA fragments to establish whether the cloned fragments in the probe mixture overlap and are therefore contiguous to the immobilized clones on the array. For example, Lehrach, *et al.*, describe such a process.

The arrays of immobilized DNA fragments may also be used for genetic diagnostics. To illustrate, an array containing multiple forms of a mutated gene or genes can be probed with a labeled mixture of a

patient's DNA which will preferentially interact with only one of the immobilized versions of the gene.

The detection of this interaction can lead to a medical diagnosis. Arrays of immobilized DNA fragments can also be used in DNA probe diagnostics. For example, the identity of a pathogenic microorganism can be established unambiguously by hybridizing a sample of the unknown pathogen's DNA to an array containing many types of known pathogenic DNA. A similar technique can also be used for unambiguous genotyping of any organism. Other molecules of genetic interest, such as cDNA's and RNA's can be immobilized on the array or alternately used as the labeled probe mixture that is applied to the array.

In one application, an array of cDNA clones representing genes is hybridized with total cDNA from an organism to monitor gene expression for research or diagnostic purposes. Labeling total cDNA from a normal cell with one color fluorophore and total cDNA from a diseased cell with another color fluorophore and simultaneously hybridizing the two cDNA samples to the same array of cDNA clones allows for differential gene expression to be measured as the ratio of the two fluorophore intensities. This two-color experiment can be used to monitor gene expression in different tissue types, disease states, response to drugs, or response to environmental factors. & An example of this approach is illustrated in Examples 2, described with respect to Fig. 8.

By way of example and without implying a limitation of scope, such a procedure could be used to simultaneously screen many patients against all known mutations in a disease gene. This invention could be used in the form of, for example, 96 identical 0.9 cm × 2.2 cm microarrays fabricated on a single 12 cm × 18 cm

sheet of plastic-backed nitrocellulose where each microarray could contain, for example, 100 DNA fragments representing all known mutations of a given gene. The region of interest from each of the DNA samples from 96 patients could be amplified, labeled, and hybridized to the 96 individual arrays with each assay performed in 100 microliters of hybridization solution. The approximately 1 thick silicone rubber barrier elements between individual arrays prevent cross contamination of the patient samples by sealing the pores of the nitrocellulose and by acting as a physical barrier between each microarray. The solid support containing all 96 microarrays assayed with the 96 patient samples is incubated, rinsed, detected and analyzed as a single sheet of material using standard radioactive, fluorescent, or colorimetric detection means (Maniatas, *et al.*, 1989). Previously, such a procedure would involve the handling, processing and tracking of 96 separate membranes in 96 separate sealed chambers. By processing all 96 arrays as a single sheet of material, significant time and cost savings are possible.

The assay format can be reversed where the patient or organism's DNA is immobilized as the array elements and each array is hybridized with a different mutated allele or genetic marker. The gridded solid support can also be used for parallel non-DNA ELISA assays. Furthermore, the invention allows for the use of all standard detection methods without the need to remove the shallow barrier elements to carry out the detection step.

In addition to the genetic applications listed above, arrays of whole cells, peptides, enzymes, antibodies, antigens, receptors, ligands, phospholipids, polymers, drug cogener preparations or

chemical substances can be fabricated by the means described in this invention for large scale screening assays in medical diagnostics, drug discovery, molecular biology, immunology and toxicology.

The multi-cell substrate aspect of the invention allows for the rapid and convenient screening of many DNA probes against many ordered arrays of DNA fragments. This eliminates the need to handle and detect many individual arrays for performing mass screenings for genetic research and diagnostic applications. Numerous microarrays can be fabricated on the same solid support and each microarray reacted with a different DNA probe while the solid support is processed as a single sheet of material.

The following examples illustrate, but in no way are intended to limit, the present invention.

## Example 1

### Genomic-Complexity Hybridization to Micro DNA Arrays Representing the Yeast *Saccharomyces cerevisiae* Genome with Two-Color Fluorescent Detection

The array elements were randomly amplified PCR (Bohlander, *et al.*, 1992) products using physically mapped lambda clones of *S. cerevisiae* genomic DNA templates (Riles, *et al.*, 1993). The PCR was performed directly on the lambda phage lysates resulting in an amplification of both the 35 kb lambda vector and the 5-15 kb yeast insert sequences in the form of a uniform distribution of PCR product between 250-1500 base pairs in length. The PCR product was purified using Sephadex G50 gel filtration (Pharmacia, Piscataway, NJ) and concentrated by evaporation to dryness at room temperature overnight. Each of the 864 amplified

lambda clones was rehydrated in 15 μl of 3 × SSC in preparation for spotting onto the glass.

The micro arrays were fabricated on microscope slides which were coated with a layer of poly-l-lysine (Sigma). The automated apparatus described in Section IV loaded 1 μl of the concentrated lambda clone PCR product in 3 × SSC directly from 96 well storage plates into the open capillary printing element and deposited ~5 nl of sample per slide at 380 micron spacing between spots, on each of 40 slides. The process was repeated for all 864 samples and 8 control spots. After the spotting operation was complete, the slides were rehydrated in a humid chamber for 2 hours, baked in a dry 80° vacuum oven for 2 hours, rinsed to remove un-absorbed DNA and then treated with succinic anhydride to reduce non-specific adsorption of the labeled hybridization probe to the poly-l-lysine coated glass surface. Immediately prior to use, the immobilized DNA on the array was denatured in distilled water at 90° for 2 minutes.

For the pooled chromosome experiment, the 16 chromosomes of *Saccharomyces cerevisiae* were separated in a CHEF agarose gel apparatus (Biorad, Richmond, CA). The six largest chromosomes were isolated in one gel slice and the smallest 10 chromosomes in a second gel slice. The DNA was recovered using a gel extraction kit (Qiagen, Chatsworth, CA). The two chromosome pools were randomly amplified in a manner similar to that used for the target lambda clones. Following amplification, 5 micrograms of each of the amplified chromosome pools were separately random-primer labeled using Klenow polymerase (Amersham, Arlington Heights, IL) with a lissamine conjugated nucleotide analog (Dupont NEN, Boston, MA) for the pool containing the six largest chromosomes, and with a fluorescein

conjugated nucleotide analog (BMB) for the pool containing smallest ten chromosomes. The two pools were mixed and concentrated using an ultrafiltration device (Amicon, Danvers, MA).

Five micrograms of the hybridization probe consisting of both chromosome pools in 7.5 μl of TE was denatured in a boiling water bath and then snap cooled on ice. 2.5 μl of concentrated hybridization solution (5 × SSC and 0.1% SDS) was added and all 10 μl transferred to the array surface, covered with a cover slip, placed in a custom-built single-slide humidity chamber and incubated at 60° for 12 hours. The slides were then rinsed at room temperature in 0.1 × SSC and 0.1%SDS for 5 minutes, cover slipped and scanned.

A custom built laser fluorescent scanner was used to detect the two-color hybridization signals from the 1.8 × 1.8 cm array at 20 micron resolution. The scanned image was gridded and analyzed using custom image analysis software. After correcting for optical crosstalk between the fluorophores due to their overlapping emission spectra, the red and green hybridization values for each clone on the array were correlated to the known physical map position of the clone resulting in a computer-generated color karyotype of the yeast genome.

Figure 6 shows the hybridization pattern of the two chromosome pools. A red signal indicates that the lambda clone on the array surface contains a cloned genomic DNA segment from one of the largest six yeast chromosomes. A green signal indicates that the lambda clone insert comes from one of the smallest ten yeast chromosomes. Orange signals indicate repetitive sequences which cross hybridized to both chromosome pools. Control spots on the array confirm that the hybridization is specific and reproducible.

The physical map locations of the genomic DNA fragments contained in each of the clones used as array elements have been previously determined by Olson and co-workers (Riles, *et al.*) allowing for the automatic generation of the color karyotype shown in Figure 7. The color of a chromosomal section on the karyotype corresponds to the color of the array element containing the clone from that section. The black regions of the karyotype represent false negative dark spots on the array (10%) or regions of the genome not covered by the Olson clone library (90%). Note that the largest six chromosomes are mainly red while the smallest ten chromosomes are mainly green matching the original CHEF gel isolation of the hybridization probe. Areas of the red chromosomes containing green spots and vice-versa are probably due to spurious sample tracking errors in the formation of the original library and in the amplification and spotting procedures.

The yeast genome arrays have also been probed with individual clones or pools of clones that are fluorescently labeled for physical mapping purposes. The hybridization signals of these clones to the array were translated into a position on the physical map of yeast.

## Example 2

## Total cDNA Hybridized to Micro Arrays of cDNA Clones with Two-Color Fluorescent Detection

24 clones containing cDNA inserts from the plant *Arabidopsis* were amplified using PCR. Salt was added to the purified PCR products to a final concentration of 3 × SSC. The cDNA clones were spotted on poly-l-lysine coated microscope slides in a manner similar to Example 1. Among the cDNA clones was a clone

representing a transcription factor HAT 4, which had previously been used to create a transgenic line of the plant *Arabidopsis*, in which this gene is present at ten times the level found in wild-type *Arabidopsis* (Schena, *et al.*, 1992).

Total poly-A mRNA from wild type *Arabidopsis* was isolated using standard methods (Maniatis, *et al.*, 1989) and reverse transcribed into total cDNA, using fluorescein nucleotide analog to label the cDNA product (green fluorescence). A similar procedure was performed with the transgenic line of *Arabidopsis* where the transcription factor HAT4 was inserted into the genome using standard gene transfer protocols. cDNA copies of mRNA from the transgenic plant are labeled with a lissamine nucleotide analog (red fluorescence). Two micrograms of the cDNA products from each type of plant were pooled together and hybridized to the cDNA clone array in a 10 microliter hybridization reaction in a manner similar to Example 1. Rinsing and detection of hybridization was also performed in a manner similar to Example 1. Fig. 8 show the resulting hybridization pattern of the array.

Genes equally expressed in wild type and the transgenic *Arabidopsis* appeared yellow due to equal contributions of the green and red fluorescence to the final signal. The dots are different intensities of yellow indicating various levels of gene expression. The cDNA clone representing the transcription factor HAT4, expressed in the transgenic line of *Arabidopsis* but not detectably expressed in wild type *Arabidopsis*, appears as a red dot (with the arrow pointing to it), indicating the preferential expression of the transcription factor in the red-labeled transgenic *Arabidopsis* and the relative lack of expression of the

transcription factor in the green-labeled wild type *Arabidopsis*.

An advantage of the microarray hybridization format for gene expression studies is the high partial concentration of each cDNA species achievable in the 10 microliter hybridization reaction. This high partial concentration allows for detection of rare transcripts without the need for PCR amplification of the hybridization probe which may bias the true genetic representation of each discrete cDNA species.

Gene expression studies such as these can be used for genomics research to discover which genes are expressed in which cell types, disease states, development states or environmental conditions. Gene expression studies can also be used for diagnosis of disease by empirically correlating gene expression patterns to disease states.

## Example 3

## Multiplexed Colorimetric Hybridization on a Gridded Solid Support

A sheet of plastic-backed nitrocellulose was gridded with barrier elements made from silicone rubber according to the description in Section IV-A. The sheet was soaked in 10 × SSC and allowed to dry. As shown in Fig. 12, 192 M13 clones each with a different yeast inserts were arrayed 400 microns apart in four quadrants of the solid support using the automated device described in Section III. The bottom left quadrant served as a negative control for hybridization while each of the other three quadrants was hybridized simultaneously with a different oligonucleotide using the open-face hybridization technology described in Section IV-A. The first two and last four elements of

each array are positive controls for the colorimetric detection step.

The oligonucleotides were labeled with fluorescein which was detected using an anti-fluorescein antibody conjugated to alkaline phosphatase that precipitated an NBT/BCIP dye on the solid support (Amersham). Perfect matches between the labeled oligos and the M13 clones resulted in dark spots visible to the naked eye and detected using an optical scanner (HP ScanJet II) attached to a personal computer. The hybridization patterns are different in every quadrant indicating that each oligo found several unique M13 clones from among the 192 with a perfect sequence match. Note that the open capillary printing tip leaves detectable dimples on the nitrocellulose which can be used to automatically align and analyze the images.

Although the invention has been described with respect to specific embodiments and methods, it will be clear that various changes and modification may be made without departing from the invention.

IT IS CLAIMED:

1. A method of forming a microarray of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent, said method comprising,

(a) loading a solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel (i) formed by spaced-apart, coextensive elongate members, (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus,

(b) tapping the tip of the dispensing device against a solid support at a defined position on the surface, with an impulse effective to break the meniscus in the capillary channel and deposit a selected volume of solution on the surface, and

(c) repeating steps (a) and (b) until said array is formed.

2. The method of claim 1, wherein said tapping is carried out with an impulse effective to deposit a selected volume in the volume range between 0.01 to 100 nl.

3. The method of claim 1, wherein said channel is formed by a pair of spaced-apart tapered elements.

4. The method of claim 1, for forming a plurality of such arrays, wherein step (b) is applied to a selected position on each of a plurality of solid supports at each repeat cycle proceeding step (c).

5. The method of claim 1, which further includes, after performing steps (a) and (b) at least one time, reloading the reagent-dispensing device with a new reagent solution by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

6. Automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected, analyte-specific reagent, said apparatus comprising

(a) a holder for holding, at known positions, a plurality of planar supports,

(b) a reagent dispensing device having an open capillary channel (i) formed by spaced-apart, coextensive elongate members (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus,

(c) positioning means for positioning the dispensing device at a selected array position with respect to a support in said holder,

(d) dispensing means for moving the device into tapping engagement against a support with a selected impulse, when the device is positioned at a defined array position with respect to that support, with an impulse effective to break the meniscus of liquid in the capillary channel and deposit a selected volume of solution on the surface, and

(e) control means for controlling said positioning and dispensing means.

7. The apparatus of claim 6, wherein said dispensing means is effective to move said dispensing device against a support with an impulse effective to deposit a selected volume in the volume range between 0.01 to 100 nl.

8. The apparatus of claim 6, wherein said channel is formed by a pair of spaced-apart tapered elements.

9. The apparatus of claim 6, wherein the control means operates to (i) place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and (iii) dispense the reagent at a defined array position on each of the supports on said holder.

10. The apparatus of claim 6, wherein the control device further operates, at the end of a dispensing cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) remove the wash fluid prior to loading the dispensing device with a fresh selected reagent.

11. The apparatus of claim 6, wherein said device is one of a plurality of such devices which are carried on the arm for dispensing different analyte assay reagents at selected spaced array positions.

12. A substrate with a surface having a microarray of at least $10^3$ distinct polynucleotide or polypeptide biopolymers per 1 $cm^2$ surface area, each

distinct biopolymer sample (i) being disposed at a separate, defined position in said array, (ii) having a length of at least 50 subunits, and (iii) being present in a defined amount between about 0.1 femtomole and 100 nanomoles.

13. The substrate of claim 12, wherein said surface is glass slide coated with polylysine, and said biopolymers are polynucleotides.

14. The substrate of claim 12, wherein said substrate has a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film, where said grid (i) is composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film, and (ii) partitions the film into a plurality of water-impervious cells, where each cell contains such a biopolymer array.

15. A substrate with a surface array of sample-receiving cells, comprising

a water-impermeable backing,

a water-permeable film formed on the backing, and a grid formed on the film, said grid being composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film.

16. The substrate of claim 15, wherein the cells of the array each contain an array of biopolymers.

17. A substrate for use in detecting binding of labeled biopolymers to one or more of a plurality distinct polynucleotides, comprising

a non-porous, glass substrate,

a coating of a cationic polymer on said substrate, and

an array of distinct polynucleotides to said coating, where each biopolymer is disposed at a separate, defined position in a surface array of biopolymers.

18. A method of detecting differential expression of each of a plurality of genes in a first cell type with respect to expression of the same genes in a second cell types, said method comprising

producing fluorescence-labeled cDNA's from mRNA's isolated from the two cells types, where the cDNA's from the first and second cells are labeled with first and second different fluorescent reporters,

adding a mixture of the labeled cDNA's from the two cell types to an array of polynucleotides representing a plurality of known genes derived from the two cell types, under conditions that result in hybridization of the cDNA's to complementary-sequence polynucleotides in the array; and

examining the array by fluorescence under fluorescence excitation conditions in which (i) polynucleotides in the array that are hybridized predominantly to cDNA's derived from one of the first and second cell types give a distinct first or second fluorescence emission color, respectively, and (ii) polynucleotides in the array that are hybridized to substantially equal numbers of cDNA's derived from the first and second cell types give a distinct combined fluorescence emission color, respectively,

wherein the relative expression of known genes in the two cell types can be determined by the observed fluorescence emission color of each spot.

19. The method of claim 18, wherein the array of polynucleotides is formed on a substrate with a surface having an array of at least $10^2$ distinct polynucleotide or polypeptide biopolymers in a surface area of less than about 1 $cm^2$, each distinct biopolymer (i) being disposed at a separate, defined position in said array, (ii) having a length of at least 50 subunits, and (iii) being present in a defined amount between about .1 femtomole and 100 nmoles.

20. The method of claim 19, wherein said surface is a glass slide coated with polylysine, and said biopolymers are polynucleotides non-covalently bound to said polylysine.

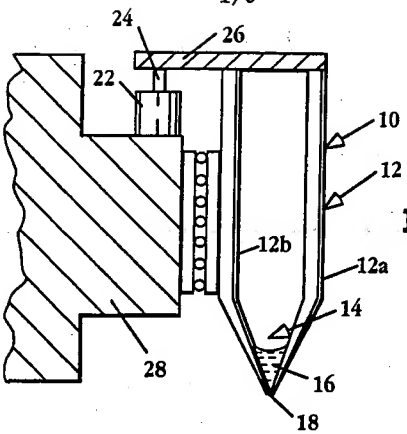



Fig. 1

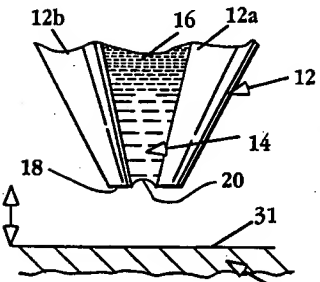



Fig. 2A

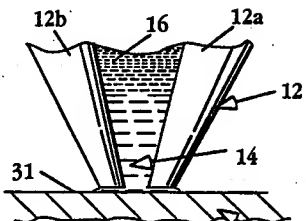



Fig. 2B

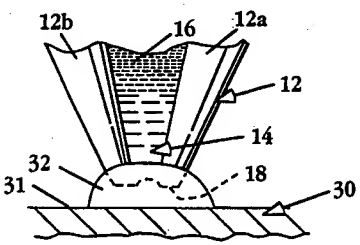



Fig. 2C

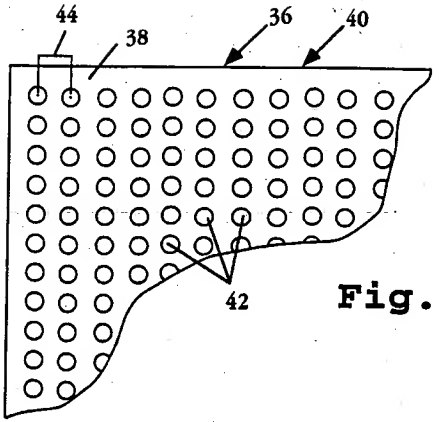

44
38
36
40
42

Fig. 3

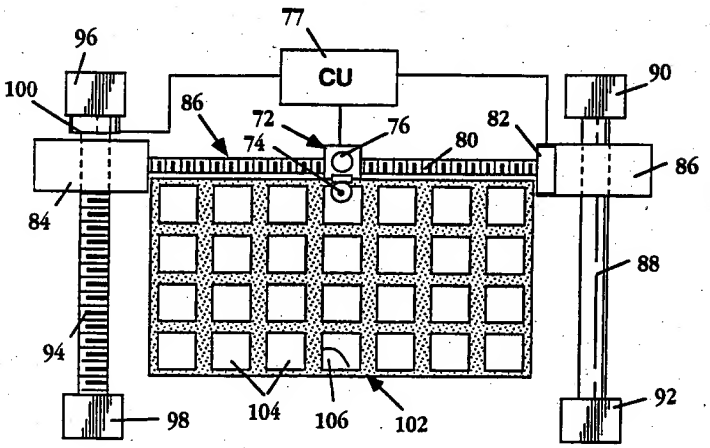

77
CU
96
100
86
72
74
76
80
82
90
86
84
88
94
98
104
106
102
92

Fig. 4

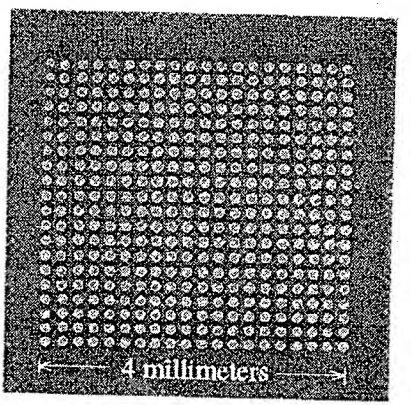



Fig. 5

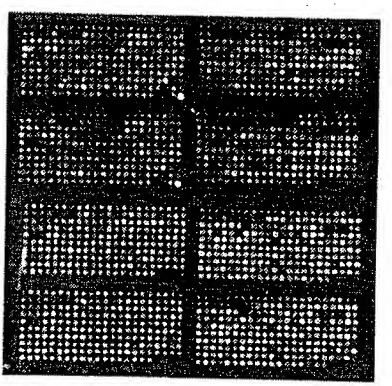


Fig. 6

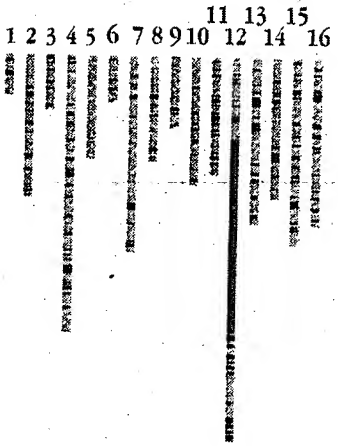



Fig. 7

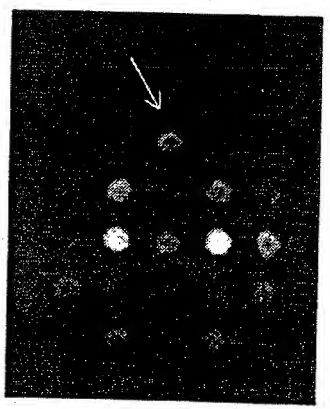


Fig. 8

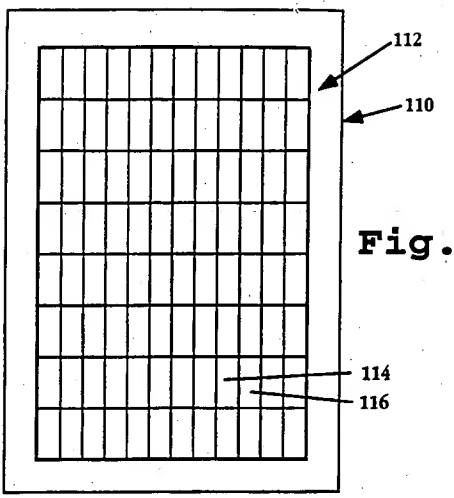



Fig. 9

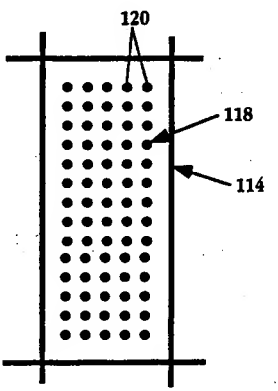



Fig. 10

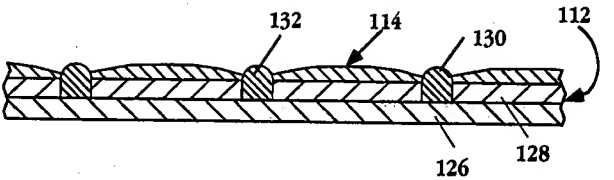



Fig. 11

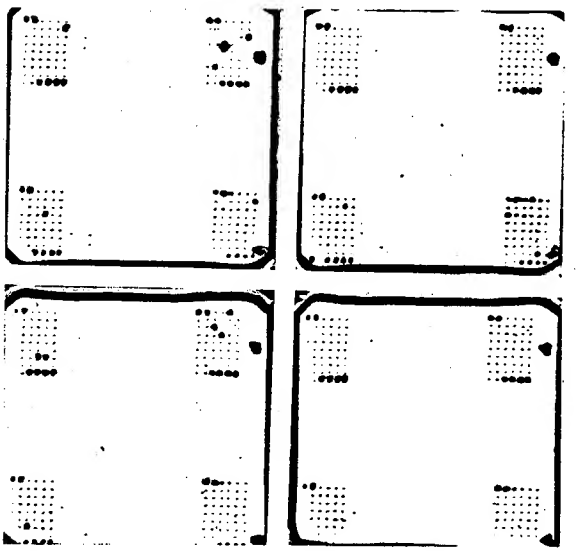


Fig. 12

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/07659

A. CLASSIFICATION OF SUBJECT MATTER
IPC(6) :G01N 33/543, 33/68
US CL :435/6; 436/518
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 422/57; 435/4,6,973; 436/518,524,527,531,805,809

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A,P | US, A, 5,338,688 (DEEG ET AL) 16 August 1994, see entire document | 1-17 |
| A | US, A, 5,204,268 (MATSUMOTO) 20 April 1993, see entire document. | 6-11 |
| A | US, A, 4,071,315 (CHATEAU) 31 January 1978, see entire document. | 12-17 |
| A | US, A, 5,100,777 (CHANG) 31 March 1992, see entire document. | 12-17 |
| A | US, A, 5,200,312 (OPRANDY) 06 April 1993, see entire document. | 12-17 |

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
15 SEPTEMBER 1995

Date of mailing of the international search report
06 OCT 1995

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Authorized officer
CHRISTOPHER CHIN

Facsimile No. (703) 305-3230

Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

*Proc. Natl. Acad. Sci. USA*
Vol. 94, pp. 2150–2155, March 1997
Biochemistry



# Discovery and analysis of inflammatory disease-related genes using cDNA microarrays

(inflammation/human genome analysis/gene discovery)

RENU A. HELLER*†, MARK SCHENA*, ANDREW CHAI*, DARI SHALON‡, TOD BEDILION‡, JAMES GILMORE‡, DAVID E. WOOLLEY§, AND RONALD W. DAVIS*

*Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305; ‡Synteni, Palo Alto, CA 94306; and §Department of Medicine, Manchester Royal Infirmary, Manchester, United Kingdom

*Contributed by Ronald W. Davis, December 27, 1996*

**ABSTRACT** **cDNA microarray technology is used to profile complex diseases and discover novel disease-related genes. In inflammatory disease such as rheumatoid arthritis, expression patterns of diverse cell types contribute to the pathology. We have monitored gene expression in this disease state with a microarray of selected human genes of probable significance in inflammation as well as with genes expressed in peripheral human blood cells. Messenger RNA from cultured macrophages, chondrocyte cell lines, primary chondrocytes, and synoviocytes provided expression profiles for the selected cytokines, chemokines, DNA binding proteins, and matrix-degrading metalloproteinases. Comparisons between tissue samples of rheumatoid arthritis and inflammatory bowel disease verified the involvement of many genes and revealed novel participation of the cytokine interleukin 3, chemokine Groα and the metalloproteinase matrix metallo-elastase in both diseases. From the peripheral blood library, tissue inhibitor of metalloproteinase 1, ferritin light chain, and manganese superoxide dismutase genes were identified as expressed differentially in rheumatoid arthritis compared with inflammatory bowel disease. These results successfully demonstrate the use of the cDNA microarray system as a general approach for dissecting human diseases.**

The recently described cDNA microarray or DNA-chip technology allows expression monitoring of hundreds and thousands of genes simultaneously and provides a format for identifying genes as well as changes in their activity (1, 2). Using this technology, two-color fluorescence patterns of differential gene expression in the root versus the shoot tissue of *Arabidopsis* were obtained in a specific array of 48 genes (1). In another study using a 1000 gene array from a human peripheral blood library, novel genes expressed by T cells were identified upon heat shock and protein kinase C activation (3).

The technology uses cDNA sequences or cDNA inserts of a library for PCR amplification that are arrayed on a glass slide with high speed robotics at a density of 1000 cDNA sequences per $cm^2$. These microarrays serve as gene targets for hybridization to cDNA probes prepared from RNA samples of cells or tissues. A two-color fluorescence labeling technique is used in the preparation of the cDNA probes such that a simultaneous hybridization but separate detection of signals provides the comparative analysis and the relative abundance of specific genes expressed (1, 2). Microarrays can be constructed from specific cDNA clones of interest, a cDNA library, or a select number of open reading frames from a genome sequencing database to allow a large-scale functional analysis of expressed sequences.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.


0027-8424/97/942150-6$2.00/0
PNAS is available online at **http://www.pnas.org.**

Because of the wide spectrum of genes and endogenous mediators involved, the microarray technology is well suited for analyzing chronic diseases. In rheumatoid arthritis (RA), inflammation of the joint is caused by the gene products of many different cell types present in the synovium and cartilage tissues plus those infiltrating from the circulating blood. The autoimmune and inflammatory nature of the disease is a cumulative result of genetic susceptibility factors and multiple responses, paracrine and autocrine in nature, from macrophages, T cells, plasma cells, neutrophils, synovial fibroblasts, chondrocytes, etc. Growth factors, inflammatory cytokines (4), and the chemokines (5) are the important mediators of this inflammatory process. The ensuing destruction of the cartilage and bone by the invading synovial tissue includes the actions of prostaglandins and leukotrienes (6), and the matrix degrading metalloproteinases (MMPs). The MMPs are an important class of Zn-dependent metallo-endoproteinases that can collectively degrade the proteoglycan and collagen components of the connective tissue matrix (7).

This paper presents a study in which the involvement of select classes of molecules in RA was examined. Also investigated were 1000 human genes randomly selected from a peripheral human blood cell library. Their differential and quantitative expression analysis in cells of the joint tissue, in diseased RA tissue and in inflammatory bowel disease (IBD) tissues was conducted to demonstrate the utility of the microarray method to analyze complex diseases by their pattern of gene expression. Such a survey provides insight not only into the underlying cause of the pathology, but also provides the opportunity to selectively target genes for disease intervention by appropriate drug development and gene therapies.

## METHODS

**Microarray Design, Development, and Preparation.** Two approaches for the fabrication of cDNA microarrays were used in this study. In the first approach, known human genes of probable significance in RA were identified. Regions of the clones, preferably 1 kb in length, were selected by their proximity to the 3′ end of the cDNA and for areas of least identity to related and repetitive sequences. Primers were synthesized to amplify the target regions by standard PCR protocols (3). Products were

Abbreviations: RA, rheumatoid arthritis; MMP, matrix-degrading metalloproteinase; IBD, inflammatory bowel disease; LPS, lipopolysaccharide; PMA, phorbol 12-myristate 13-acetate; TNF-α, tumor necrosis factor α; IL, interleukin; TGF-β, transforming growth factor β; GCSF, granulocyte colony-stimulating factor; MIP, macrophage inflammatory protein; MIF, migration inhibitory factor; HME, human matrix metallo-elastase; RANTES, regulated upon activation, normal T cell expressed and secreted; Gel, gelatinase; VCAM, vascular cell adhesion molecule; ICE, IL-1 converting enzyme; PUMP, putative metalloproteinase; MnSOD, manganese superoxide dismutase; TIMP, tissue inhibitor of metalloproteinase; MCP, macrophage chemotactic protein.

†To whom reprint requests should be sent at the present address: Roche Bioscience, S3–1, 3401 Hillview Avenue, Palo Alto, CA 94304.

verified by gel electrophoresis and purified with Qiaquick 96-well purification kit (Qiagen, Chatsworth, CA), lyophilized (Savant), and resuspended in 5 $\mu$l of 3× standard saline citrate (SSC) buffer for arraying. In the second approach, the microarray containing the 1056 human genes from the peripheral blood lymphocyte library was prepared as described (3).

**Tissue Specimens.** Rheumatoid synovial tissue was obtained from patients with late stage classic RA undergoing remedial synovectomy or arthroplasty of the knee. Synovial tissue was separated from any associated connective tissue or fat. One gram of each synovial specimen was subjected to RNA extraction within 40 min of surgical excision, or explants were cultured in serum-free medium to examine any changes under *in vitro* conditions. For IBD, specimens of macroscopically inflamed lower intestinal mucosa were obtained from patients with Crohn disease undergoing remedial surgery. The hypertrophied mucosal tissue was separated from underlying connective tissue and extracted for RNA.

**Cultured Cells.** The Mono Mac-6 (MM6) monocytic cells (8) were grown in RPMI medium. Human chondrosarcoma SW1353 cells, primary human chondrocytes, and synoviocytes (9, 10) were cultured in DMEM; all culture media were supplemented with 10% fetal bovine serum, 100 $\mu$g/ml streptomycin, and 500 units/ml penicillin. Treatment of cells with lipopolysaccharide (LPS) endotoxin at 30 ng/ml, phorbol 12-myristate 13-acetate (PMA) at 50 ng/ml, tumor necrosis factor $\alpha$ (TNF-$\alpha$) at 50 ng/ml, interleukin (IL)-1$\beta$ at 30 ng/ml, or transforming growth factor-$\beta$ (TGF-$\beta$) at 100 ng/ml is described in the figure legends.

**Fluorescent Probe, Hybridization, and Scanning.** Isolation of mRNA, probe preparation, and quantitation with *Arabidopsis* control mRNAs was essentially as described (3) except for the following minor modification. Following the reverse transcriptase step, the appropriate Cy3- and Cy5-labeled samples were pooled; mRNA degraded by heating the sample to 65°C for 10 min with the addition of 5 $\mu$l of 0.5M NaOH plus 0.5 ml of 10 mM EDTA. The pooled cDNA was purified from unincorporated nucleotides by gel filtration in Centri-spin columns (Princeton Separations, Adelphia, NJ). Samples were lyophilized and dissolved in 6 $\mu$l of hybridization buffer (5× SSC plus 0.2% SDS). Hybridizations, washes, scanning, quantitation procedures, and pseudocolor representations of fluorescent images have been described (3). Scans for the two fluorescent probes were normalized either to the fluorescence intensity of *Arabidopsis* mRNAs spiked into the labeling reactions (see Figs. 2–4) or to the signal intensity of $\beta$-actin and glyceraldehyde-3-phosphate dehydrogenase (GAPDH; see Fig. 5).

## RESULTS

**Ninety-Six-Gene Microarray Design.** The actions of cytokines, growth factors, chemokines, transcription factors, MMPs, prostaglandins, and leukotrienes are well recognized in inflammatory disease, particularly RA (11–14). Fig. 1 displays the selected genes for this study and also includes control cDNAs of housekeeping genes such as $\beta$-actin and GAPDH and genes from *Arabidopsis* for signal normalization and quantitation (row A, columns 1–12).

**Defining Microarray Assay Conditions.** Different lengths and concentrations of target DNA were tested by arraying PCR-

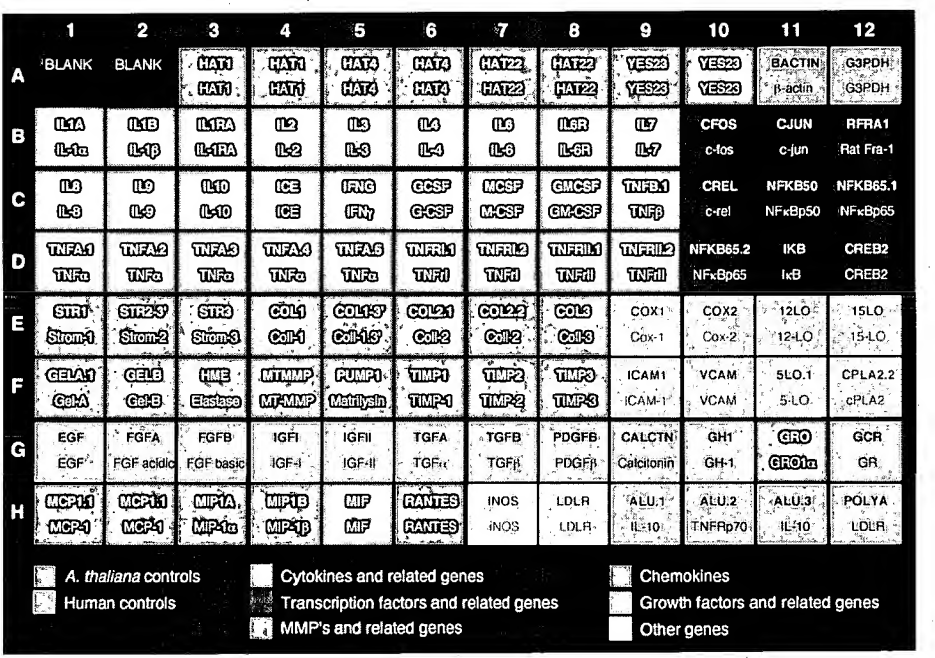



FIG. 1. Ninety-six-element microarray design. The target element name and the corresponding gene are shown in the layout. Some genes have more than one target element to guarantee specificity of signal. For TNF the targets represent decreasing lengths of 1, 0.8, 0.6, 0.4, and 0.2 kb from left to right.

amplified products ranging from 0.2 to 1.2 kb at concentrations of 1 $\mu g/\mu l$ or less. No significant difference in the signal levels was observed within this range of target size and only with 0.2-kb length was a signal reduced upon an 8-fold dilution of the 1 $\mu g/\mu l$ sample (data not shown). In this study the average length of the targets was 1 kb, with a few exceptions in the range of ≈300 bp, arrayed at a concentration of 1 $\mu g/\mu l$. Normally one PCR provided sufficient material to fabricate up to 1000 microarray targets.

In considering positional effects in the development of the targets for the microarrays, selection was biased toward the 3′ proximal regions, because the signal was reduced if the target fragment was biased toward the 5′ end (data not shown). This result was anticipated since the hybridizing probe is prepared by reverse transcription with oligo(dT)-primed mRNA and is richer in 3′ proximal sequences. Cross-hybridizations of probes to targets of a gene family were analyzed with the matrix metalloproteinases as the example because they can show regions of sequence identities of greater than 70%. With collagenase-1 (Col-1) and collagenase-2 (Col-2) genes as targets with up to 70% sequence identity, and stromelysin-1 (Strom-1) and stromelysin-2 (Strom-2) genes with different degrees of identity, our results showed that a short region of overlap, even with 70–90% sequence identity, produced a low level of cross-hybridization. However, shorter regions of identity spread over the length of the target resulted in cross-hybridization (data not shown). For closely related genes, targets were designed by avoiding long stretches of homology. For members of a gene family two or more target regions were included to discriminate between specificity of signal versus cross-hybridization.

**Monitoring Differential Expression in Cultured Cell Lines.** In RA tissue, the monocyte/macrophage population plays a prominent role in phagocytic and immunomodulatory activities. Typ-

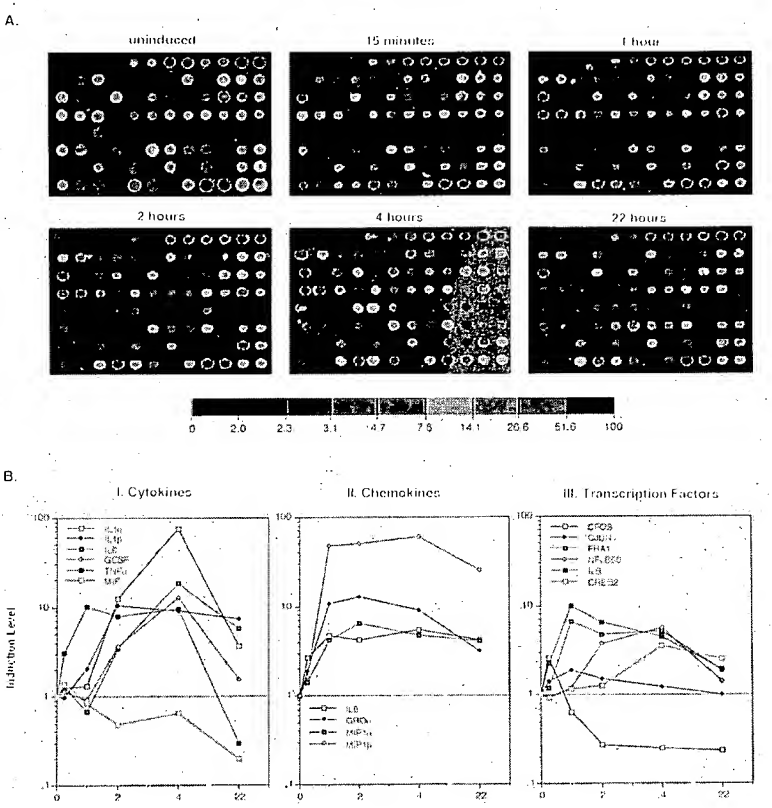



FIG. 2. Time course for LPS/PMA-induced MM6 cells. Array elements are described in Fig. 1. (*A*) Pseudocolor representations of fluorescent scans correspond to gene expression levels at each time point. The array is made up of 8 *Arabidopsis* control targets and 86 human cDNA targets, the majority of which are genes with known or suspected involvement in inflammation. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation. Fluorescent probes were made by labeling mRNA from untreated MM6 cells or LPS and PMA treated cells. mRNA was isolated at indicated times after induction. (*B I–III*) The two-color samples were cohybridized, and microarray scans provided the data for the levels of select transcripts at different time points relative to abundance at time zero. The analysis was performed using normalized data collected from 8-bit images.

ically these cells, when triggered by an immunogen, produce the proinflammatroy cytokines TNF and IL-1. We have used the monocyte cell line MM6 and monitored changes in gene expression upon activation with LPS endotoxin, a component of Gram-negative bacterial membranes, and PMA, which augments the action of LPS on TNF production (15). RNA was isolated at different times after induction and used for cDNA probe preparation. From this time course it was clear that TNF expression was induced within 15 min of treatment, reached maximum levels in 1 hr, remained high until 4 hr and subsequently declined (Fig. 2*A*). Many other cytokine genes were also transiently activated, such as IL-1$\alpha$ and -$\beta$, IL-6, and granulocyte colony-stimulating factor (GCSF). Prominent chemokines activated were IL-8, macrophage inflammatory protein (MIP)-1$\beta$, more so than MIP-1$\alpha$, and Gro$\alpha$ or melanoma growth stimulatory factor. Migration inhibitory factor (MIF) expressed in the uninduced state declined in LPS-activated cells. Of the immediate early genes, the noticeable ones were *c-fos*, *fra-1*, *c-jun*, NF-$\kappa$Bp50, and I$\kappa$B, with *c-rel* expression observed even in the uninduced state (Fig. 2*B*). These expression patterns are consistent with reported patterns of activation of certain LPS- and PMA-induced genes (12). Demonstrated here is the unique ability of this system to allow parallel visualization of a large number of gene activities over a period of time.

SW1353 cells is a line derived from malignant tumors of the cartilage and behaves much like the chondrocytes upon stimulation with TNF and IL-1 in the expression of MMPs (9). In addition to confirming our earlier observations with Northern blots on Strom-1, Col-1, and Col-3 expression (9), gelatinase (Gel) A, putative metalloproteinase (PUMP)-1 membrane-type matrix metalloproteinase, tissue inhibitors of matrix metalloproteinases or tissue inhibitor of metalloproteinase 1 (TIMP-1); -2, and -3 were also expressed by these cells together with the human matrix metallo-elastase (HME; Fig. 3*A*). HME induction was estimated to be ≈50-fold and was greater than any of the other MMPs examined (Fig. 3*B*). This result was unexpected because HME is reportedly expressed only by alveolar macrophage and placental cells (16). Expression of the cytokines and chemokines, IL-6, IL-8, MIF, and MIP-1$\beta$ was also noted. A variety of other genes, including certain transcription factors, were also up-regulated (Fig. 3), but the overall time-dependent expression of genes in the SW1353 cells was qualitatively distinct from the MM6 cells.

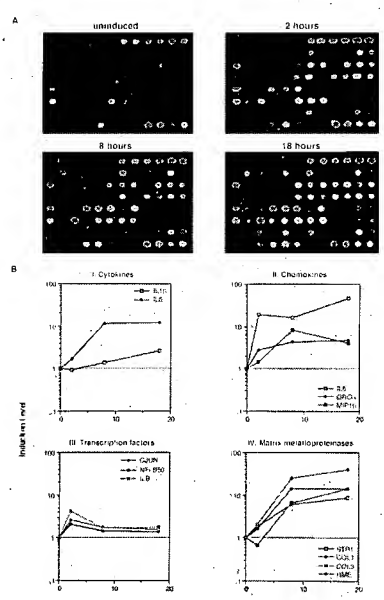



FIG. 3. Time course for IL-1$\beta$ and TNF-induced SW1353 cells using the inflammation array (Fig. 1). (*A*) Pseudocolor representation of fluorescent scans correspond to gene expression levels at each time point. (*B I–IV*) Relative levels of selected genes at different time points compared with time zero.

Quantitation of differential gene expression (Figs. 2*B* and 3*B*) was achieved with the simultaneous hybridization of Cy3-labeled cDNA from untreated cells and Cy5-labeled cDNA from treated samples. The estimated increases in expression from these microarrays for a select number of genes including IL-1$\beta$, IL-8, MIP-1$\beta$, TNF, HME, Col-1, Col-3, Strom-1, and Strom-2 were compared with data collected from dot blot analysis. Results (not shown) were in close agreement and confirmed our earlier observations on the use of the microarray method for the quantitation of gene expression (3).

**Expression Profiles in Primary Chondrocytes and Synoviocytes of Human RA Tissue.** Given the sensitivity and the specificity of this method, expression profiles of primary synoviocytes and chondrocytes from diseased tissue were examined. Without prior exposure to inducing agents, low level expression of c-*jun*, GCSF, IL-3, TNF-$\beta$, MIF, and RANTES (regulated upon activation, normal T cell expressed and secreted) was seen as well as expression of MMPs, GelA, Strom-1, Col-1, and the three TIMPs. In this case, Col-2 hybridization was considered to be nonspecific because the second Col-2 target taken from the 3′ end of the gene gave no

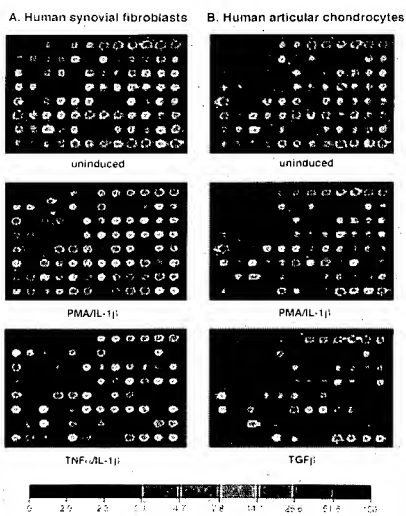



FIG. 4. Expression profiles for early passage primary synoviocytes and chondrocytes isolated from RA tissue, cultured in the presence of 10% fetal calf serum and activated with PMA and IL-1$\beta$, or TNF and IL-1$\beta$, or TGF-$\beta$ for 18 hr. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation

signal. Treatment more so with PMA and IL-1, than TNF and IL-1, produced a dramatic up-regulation in expression of several genes in both of these primary cell types. These genes are as follows: the cytokine IL-6, the chemokines IL-8 and Gro-1$\alpha$, and the MMPs; Strom-1, Col-1, Col-3, and HME; and the adhesion molecule, vascular cell adhesion molecule 1 (VCAM-1). The surprise again is HME expression in these primary cells, for reasons discussed above. From these results, the expression profiles of synoviocytes and the chondrocytes appear very similar; the differences are more quantitative than qualitative. Treatment of the primary chondrocytes with the anabolic growth factor TGF-$\beta$ had an interesting profile in that it produced a remarkable down-regulation of genes expressed in both the untreated and induced state (Fig. 4).

Given the demonstrated effectiveness of this technology, a comparative analysis of two different inflammatory disease states was conducted with probes made from RA tissue and IBD samples. RA samples were from late stage rheumatoid synovial tissue, and IBD specimens were obtained from inflamed lower intestinal mucosa of patients with Crohn disease. With both the 96-element known gene microarray and the 1000-gene microarray of cDNAs selected from a peripheral human blood cell library (3), distinct differences in gene expression patterns were evident. On the 96-gene array, RA tissue samples from different affected individuals gave similar profiles (data not shown) as did different samples from the same individual (Fig. 5). These patterns were notably similar to those observed with primary synoviocytes and chondrocytes (Fig. 4). Included in the list of prominently up-regulated genes are IL-6, the MMPs Strom-1, Col-1, GelA, HME, and in certain samples PUMP, TIMPs, particularly TIMP-1 and TIMP-3, and the adhesion molecule VCAM. Discernible levels of macrophage chemotactic protein 1 (MCP-1), MIF and RANTES were also noted. IBD samples were in comparison, rather subdued although IL-1 converting enzyme (ICE), TIMP-1, and MIF were notable in all the three different IBD samples examined here. In IBD-A, one of three individual samples, ICE, VCAM, Gro$\alpha$, and MMP expression was more pronounced than in the others.

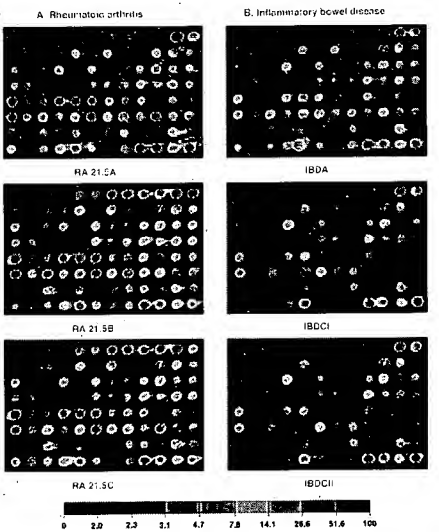



FIG. 5. Expression profiles of RA tissue (*A*) and IBD tissue (*B*). mRNA from RA tissue samples obtained from the same individual was isolated directly after excision (RA 21.5A) or maintained in culture without serum for 2 hr (RA 21.5B) or for 6 hr (RA 21.5C). Profiles from tissue samples of two other individuals (data not shown) were remarkably similar to the ones shown here. IBD-A and IBD-CI are from mRNA samples prepared directly after surgery from two separate individuals. For the IBD-CII probe, the tissue sample was cultured in medium without serum for 2 hr before mRNA preparation.

We also made use of a peripheral blood cDNA library (3) to identify genes expressed by lymphocytes infiltrating the inflamed tissues from the circulating blood. With the 1046-element array of randomly selected cDNAs from this library, probes made from RA and IBD samples showed hybridizations to a large number of genes. Of these, many were common between the two disease tissues while others were differentially expressed (data not shown). A complete survey of these genes was beyond the scope of this study, but for this report we picked three genes that were up-regulated in the RA tissue relative to IBD. These cDNAs were sequenced and identified by comparison to the GenBank database. They are TIMP-1, apoferritin light chain, and manganese superoxide dismutase (MnSOD). Differential expression of MnSOD was only observed in samples of RA tissue explants maintained in growth medium without serum for anywhere between 2 to 16 hr. These results also indicate that the expression profile of genes can be altered when explants are transferred to culture conditions.

## DISCUSSION

The speed, ease, and feasibility of simultaneously monitoring differential expression of hundreds of genes with the cDNA microarray based system (1–3) is demonstrated here in the analysis of a complex disease such as RA. Many different cell types in the RA tissue; macrophages, lymphocytes, plasma cells, neutrophils, synoviocytes, chondrocytes, etc. are known to contribute to the development of the disease with the expression of gene products known to be proinflammatory. They include the cytokines, chemokines, growth factors, MMPs, eicosanoids, and others (7, 11–14), and the design of the 96-element known gene microarray was based on this knowledge and depended on the availability of the genes. The technology was validated by confirming earlier observations on the expression of TNF by the monocyte cell line MM6, and of Col-1 and Col-3 expression in the chondrosarcoma cells and articular chondrocytes (9, 12). In our time-dependent survey the chronological order of gene activities in and between gene families was compared and the results have provided unprecedented profiles of the cytokines (TNF, IL-1, IL-6, GCSF, and MIF), chemokines (MIP-1$\alpha$, MIP-1$\beta$, IL-8, and Gro-1), certain transcription factors, and the matrix metalloproteinases (GelA, Strom-1, Col-1, Col-3, HME) in the macrophage cell line MM6 and in the SW1353 chondrosarcoma cells.

Earlier reports of cytokine production in the diseased state had established a model in which TNF is a major participant in RA. Its expression reportedly preceded that of the other cytokines and effector molecules (4). Our results strongly support these results as demonstrated in the time course of the MM6 cells where TNF induction preceded that of IL-1$\alpha$ and IL-$\beta$ followed by IL-6 and GCSF. These expression profiles demonstrate the utility of the microarrays in determining the hierarachy of signaling events.

In the SW1353 chondrosarcoma cells, all the known MMPs and TIMPs were examined simultaneously. HME expression was discovered, which previously had been observed in only the stromal cells and alveolar macrophages of smoker's lungs and in placental tissue. Its presence in cells of the RA tissue is meaningful because its activity can cause significant destruction of elastin and basement membrane components (16, 17). Expression profiles of synovial fibroblasts and articular chondrocytes were remarkably similar and not too different from the SW1353 cells, indicating that the fibroblast and the chondrocyte can play equally aggressive roles in joint erosion. Prominent genes expressed were

the MMPs, but chemokines and cytokines were also produced by these cells. The effect of the anabolic growth factor TGF-$\beta$ was profoundly evident in demonstrating the down regulation of these catabolic activities.

RA tissue samples undeniably reflected profiles similar to the cell types examined. Active genes observed were IL-3, IL-6, ICE, the MMPs including HME and TIMPs, chemokines IL-8, Groα, MIP, MIF, and RANTES, and the adhesion molecule VCAM. Of the growth factors, fibroblast growth factor $\beta$ was observed most frequently. In comparison, the expression patterns in the other inflammatory state (i.e., IBD) were not as marked as in the RA samples, at least as obtained from the tissue samples selected for this study.

As an alternative approach, the 1046 cDNA microarray of randomly selected genes from a lymphocyte library was used to identify genes expressed in RA tissue (3). Many genes on this array hybridized with probes made from both RA and IBD tissue samples. The results are not surprising because inflammatory tissue is abundantly supplied with cell types infiltrating from the circulating blood, made apparent also by the high levels of chemokine expression in RA tissue. Because of the magnitude of the effort required to identify all the hybridized genes, we have for this report chosen to describe only three differentially expressed genes mainly to verify this method of analysis.

Of the large number of genes observed here, a fair number were already known as active participants in inflammatory disease. These are TNF, IL-1, IL-6, IL-8, GCSF, RANTES, and VCAM. The novel participants not previously reported are HME, IL-3, ICE, and Groα. With our discovery of HME expression in RA, this gene becomes a target for drug intervention. ICE is a cysteine protease well known for its IL-1$\beta$ processing activity (18), and recognized for its role in apoptotic cell death (19). Its expression in RA tissue is intriguing. IL-3 is recognized for its growth-promoting activity in hematopoietic cell lineages, is a product of activated T cells (20), and its expression in synoviocytes and chondrocytes of RA tissue is a novel observation.

Like IL-8, Groα, is a C-X-C subgroup chemokine and is a potent neutrophil and basophil chemoattractant. It down-regulates the expression of types I and III interstitial collagens (21, 22) and is seen here produced by the MM6 cells, in primary synoviocytes, and in RA tissue. With the presence of RANTES, MCP, and MIP-1$\beta$, the C-C chemokines (23) migration and infiltration of monocytes, particularly T cells, into the tissue is also enhanced (5) and aid in the trafficking and recruitment of leukocytes into the RA tissue. Their activation, phagocytosis, degranulation, and respiratory bursts could be responsible for the induction of MnSOD in RA. MnSOD is also induced by TNF and IL-1 and serves a protective function against oxidative damage. The induction of the ferritin light chain encoding gene in this tissue may be for reasons similar to those for MnSOD. Ferritin is the major intracellular iron storage protein and it is responsive to intracellular oxidative stress and reactive oxygen intermediates generated during inflammation (24, 25). The active expression of TIMP-1 in RA tissue, as detected by the 1000-element array, is no surprise because our results have repeatedly shown TIMP-1 to be expressed in the constitutive and induced states of RA cells and tissues.

The suitability of the cDNA microarray technology for profiling diseases and for identifying disease related genes is well documented here. This technology could provide new targets for drug development and disease therapies, and in doing so allow for improved treatment of chronic diseases that are challenging because of their complexity.

We would like to thank the following individuals for their help in obtaining reagents or providing cDNA clones to use as templates in target preparation: N. Arai, P. Cannon, D. R. Cohen, T. Curran, V. Dixit, D. A. Geller, G. I. Goldberg, M. Karin, M. Lotz, L. Matrisian, G. Nolan, C. Lopez-Otin, T. Schall, S. Shapiro, I. Verma, and H. Van Wart. Support for R.W.D., M.S., and R.A.H. was provided by the National Institutes of Health (Grants R37HG00198 and HG00205).

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270,** 467–470.
2. Shalon, D., Smith, S. & Brown, P. O. (1996) *Genome Res.* **6,** 639–645.
3. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 10614–10619.
4. Feldmann, M., Brennan F. M. & Maini, R. N. (1996) *Rheumatoid Arthritis Cell* **85,** 307–310.
5. Schall, T. J. (1994) in *The Cytokine Handbook*, ed. Thomson, A. W. (Academic, New York), 2nd Ed., pp. 410–460.
6. Lotz, M. F., Blanco, J., Von Kempis, J., Dudler, J., Maier, R., Villiger P. M. & Geng, Y. (1995) *J. Rheumatol.* **22,** Supplement 43, 104–108.
7. Birkedal-Hansen, H., Moore, W. G. I., Bodden, M. K., Windsor, L. J., Birkedal-Hansen, B., DeCarlo, A. &. Engler, J. A. (1993) *Crit. Rev. Oral Biol. Med.* **4,** 197–250.
8. Zeigler-Heitbrock, H. W. L., Thiel, E., Futterer, A., Volker, H., Wirtz, A. & Reithmuller, G. (1988) *Int. J. Cancer* **41,** 456–461.
9. Borden, P., Solymar, D., Sucharczuk, A., Lindman, B., Cannon, P. & Heller, R. A. (1996) *J. Biol. Chem.* **271,** 23577–23581.
10. Gadher, S. J. & Woolley, D. E. (1987) *Rheumatol. Int.* **7,** 13–22.
11. Harris, E. D., Jr. (1990) *New Engl. J. Med.* **322,** 1277–1289.
12. Firestein, G. S. (1996) in *Textbook of Rheumatology*, eds. Kelly, W. N., Harris, E. D., Ruddy, S. & Sledge, C. B. (Saunders, Philadelphia), 5th Ed. pp. 5001–5047.
13. Alvaro-Garcia, J. M., Zvaifler, Nathan J., Brown, C. B., Kaushansky, K. & Firestein, Gary S. (1991) *J. Immunol.* **146,** 3365–3371.
14. Firestein, G. S., Alvaro-Grarcia, J. M. & Maki, R. (1990) *J. Immunol.* **144,** 3347–3352.
15. Pradines-Figueres, A. & Raetz, C. R. H. (1992) *J. Biol. Chem.* **267,** 23261–23268.
16. Shapiro, S. D., Kobayashi, D. L. & Ley, T. J. (1993) *J. Biol. Chem.* **208,** 23824–23829.
17. Shipley, M. J., Wesselschmidt, R. L., Kobayashi, D. K., Ley, T. J. & Shapiro, S. D. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 3042–3946.
18. Cerreti, D. P., Kozlosky, C. J., Mosley, B., Nelson, N., Van Ness, K., Greenstreet, T. A., March, C. J., Kronheim, S. R., Druck, T., Cannizaro, L. A., Huebner, K. & Black, R. A. (1992) *Science* **256,** 97–100.
19. Miura, M., Zhu, H., Rotello, R., Hartweig, E. A. & Yuan, J. (1993) *Cell* **75,** 653–660.
20. Arai, K., Lee, F., Miyajima, A., Shoichiro, M., Arai, N. & Takashi, Y. (1990) *Annu. Rev. Biochem.* **59,** 783–836.
21. Geiser, T., Dewald, B., Ehrengruber, M. U., Lewis, I. C. & Baggiolini, M. (1993) *J. Biol. Chem.* **268,** 15419–15424.
22. Unemori, E. N., Amento, E. P., Bauer, E. A. & Horuk, R. (1993) *J. Biol. Chem.* **268,** 1338–1342.
23. Robinson, E., Keystone, E. C., Schall, T. J., Gillet, N. & Fish, E. N. (1995) *Clin. Exp. Immunol.* **101,** 398–407.
24. Roeser, H. (1980) in *Iron Metabolism in Biochemistry and Medicine*, eds. Jacobs, A. & Worwood, M. (Academic, New York), Vol. 2, pp. 605–640.
25. Kwak, E. L., Larochelle, D. A., Beaumont, C., Torti, S. V. & Torti, F. M. (1995) *J. Biol. Chem.* **270,** 15285–15293.

**PCT** WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C07H 21/04 — A1

(11) International Publication Number: **WO 97/13877**

(43) International Publication Date: 17 April 1997 (17.04.97)

(21) International Application Number: PCT/US96/16342

(22) International Filing Date: 11 October 1996 (11.10.96)

(30) Priority Data:
PCT/US95/12791 12 October 1995 (12.10.95) WO
*(34) Countries for which the regional or international application was filed:* US et al.
PCT/US96/09513 6 June 1996 (06.06.96) WO
*(34) Countries for which the regional or international application was filed:* US et al.

(60) Parent Application or Grant
(63) Related by Continuation
US Not furnished (CIP)
Filed on Not furnished

(71) Applicant *(for all designated States except US)*: LYNX THERAPEUTICS, INC. [US/US]; 3832 Bay Center Place, Hayward, CA 94545 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: MARTIN, David, W. [US/US]; Lynx Therapeutics, Inc., 3832 Bay Center Place, Hayward, CA 94545 (US).

(74) Agent: POWERS, Vincent, M.; Dehlinger & Associates, Post Office Box 60850, Palo Alto, CA 94306-0850 (US).

(81) Designated States: AU, CA, CZ, EE, FI, HU, JP, KR, LT, LV, NO, NZ, PL, RU, SG, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

**Published**
*With international search report.*
*Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.*

(54) Title: MEASUREMENT OF GENE EXPRESSION PROFILES IN TOXICITY DETERMINATION

(57) Abstract

A method is provided for assessing the toxicity of a compound in a test organism by measuring gene expression profiles of selected tissues. Gene expression profiles are measured by massively parallel signature sequencing of cDNA libraries constructed from mRNA extracted from the selected tissues. Gene expression profiles provide extensive information on the effects of administering a compound to a test organism in both acute toxicity tests and in prolonged and chronic toxicity tests.

*FOR THE PURPOSES OF INFORMATION ONLY*

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

| | | | | | |
|---|---|---|---|---|---|
| AM | Armenia | GB | United Kingdom | MW | Malawi |
| AT | Austria | GE | Georgia | MX | Mexico |
| AU | Australia | GN | Guinea | NE | Niger |
| BB | Barbados | GR | Greece | NL | Netherlands |
| BE | Belgium | HU | Hungary | NO | Norway |
| BF | Burkina Faso | IE | Ireland | NZ | New Zealand |
| BG | Bulgaria | IT | Italy | PL | Poland |
| BJ | Benin | JP | Japan | PT | Portugal |
| BR | Brazil | KE | Kenya | RO | Romania |
| BY | Belarus | KG | Kyrgystan | RU | Russian Federation |
| CA | Canada | KP | Democratic People's Republic of Korea | SD | Sudan |
| CF | Central African Republic | | | SE | Sweden |
| CG | Congo | KR | Republic of Korea | SG | Singapore |
| CH | Switzerland | KZ | Kazakhstan | SI | Slovenia |
| CI | Côte d'Ivoire | LI | Liechtenstein | SK | Slovakia |
| CM | Cameroon | LK | Sri Lanka | SN | Senegal |
| CN | China | LR | Liberia | SZ | Swaziland |
| CS | Czechoslovakia | LT | Lithuania | TD | Chad |
| CZ | Czech Republic | LU | Luxembourg | TG | Togo |
| DE | Germany | LV | Latvia | TJ | Tajikistan |
| DK | Denmark | MC | Monaco | TT | Trinidad and Tobago |
| EE | Estonia | MD | Republic of Moldova | UA | Ukraine |
| ES | Spain | MG | Madagascar | UG | Uganda |
| FI | Finland | ML | Mali | US | United States of America |
| FR | France | MN | Mongolia | UZ | Uzbekistan |
| GA | Gabon | MR | Mauritania | VN | Viet Nam |

# MEASUREMENT OF GENE EXPRESSION PROFILES IN TOXICITY DETERMINATION

## Field of the Invention

The invention relates generally to methods for detecting and monitoring phenotypic changes in in vitro and in vivo systems for assessing and/or determining the toxicity of chemical compounds, and more particularly, the invention relates to a method for detecting and monitoring changes in gene expression patterns in in vitro and in vivo systems for determining the toxicity of drug candidates.

## BACKGROUND

The ability to rapidly and conveniently assess the toxicity of new compounds is extremely important. Thousands of new compounds are synthesized every year, and many are introduced to the environment through the development of new commercial products and processes, often with little knowledge of their short term and long term health effects. In the development of new drugs, the cost of assessing the safety and efficacy of candidate compounds is becoming astronomical: It is estimated that the pharmaceutical industry spends an average of about 300 million dollars to bring a new pharmaceutical compound to market, e.g. Biotechnology, 13: 226-228 (1995). A large fraction of these costs are due to the failure of candidate compounds in the later stages of the developmental process. That is, as the assessment of a candidate drug progresses from the identification of a compound as a drug candidate--for example, through relatively inexpensive binding assays or in vitro screening assays, to pharmacokinetic studies, to toxicity studies, to efficacy studies in model systems, to preliminary clinical studies, and so on, the costs of the associated tests and analyses increases tremendously. Consequently, it may cost several tens of millions of dollars to determine that a once promising candidate compound possesses a side effect or cross reactivity that renders it commercially infeasible to develop further. A great challenge of pharmaceutical development is to remove from further consideration as early as possible those compounds that are likely to fail in the later stages of drug testing.

Drug development programs are clearly structured with this objective in mind; however, rapidly escalating costs have created a need to develop even more stringent and less expensive screens in the early stages to identify false leads as soon as possible. Toxicity assessment is an area where such improvements may be made, for both drug development and for assessing the environmental, health, and safety effects of new compounds in general.

Typically the toxicity of a compound is determined by administering the compound to one or more species of test animal under controlled conditions and by monitoring the effects on a wide range of parameters. The parameters include such things as blood chemistry, weight gain or loss, a variety of behavioral patterns, muscle tone, body temperature, respiration rate, lethality, and the like, which collectively provide a measure of the state of health of the test animal. The degree of deviation of such parameters from their normal ranges gives a measure of the toxicity of a compound. Such tests may be designed to assess the acute, prolonged, or chronic toxicity of a compound. In general, acute tests involve administration of the test chemical on one occasion. The period of observation of the test animals may be as short as a few hours, although it is usually at least 24 hours and in some cases it may be as long as a week or more. In general, prolonged tests involve administration of the test chemical on multiple occasions. The test chemical may be administered one or more times each day, irregularly as when it is incorporated in the diet, at specific times such as during pregnancy, or in some cases regularly but only at weekly intervals. Also, in the prolonged test the experiment is usually conducted for not less than 90 days in the rat or mouse or a year in the dog. In contrast to the acute and prolonged types of test, the chronic toxicity tests are those in which the test chemical is administered for a substantial portion of the lifetime of the test animal. In the case of the mouse or rat, this is a period of 2 to 3 years. In the case of the dog, it is for 5 to 7 years.

Significant costs are incurred in establishing and maintaining large cohorts of test animals for such assays, especially the larger animals in chronic toxicity assays. Moreover, because of species specific effects, passing such toxicity tests does not ensure that a compound is free of toxic effects when used in humans. Such tests do, however, provide a standardized set of information for judging the safety of new compounds, and they provide a database for giving preliminary assessments of related compounds. An important area for improving toxicity determination would be the identification of new observables which are predictive of the outcome of the expensive and tedious animal assays.

In other medical fields, there has been significant interest in applying recent advances in biotechnology, particularly in DNA sequencing, to the identification and study of differentially expressed genes in healthy and diseased organisms, e.g. Adams et al, Science, 252: 1651-1656 (1991); Matsubara et al, Gene, 135: 265-274 (1993); Rosenberg et al, International patent application, PCT/US95/01863. The objectives of such applications include increasing our knowledge of disease processes, identifying genes that play important roles in the disease process, and providing diagnostic and therapeutic approaches that exploit the expressed genes or their

products. While such approaches are attractive, those based on exhaustive, or even sampled, sequencing of expressed genes are still beset by the enormous effort required: It is estimated that 30-35 thousand different genes are expressed in a typical mammalian tissue in any given state, e.g. Ausubel et al, Editors, Current Protocols, 5.8.1-5.8.4 (John Wiley & Sons, New York, 1992). Determining the sequences of even a small sample of that number of gene products is a major enterprise, requiring industrial-scale resources. Thus, the routine application of massive sequencing of expressed genes is still beyond current commercial technology.

The availability of new assays for assessing the toxicity of compounds, such as candidate drugs, that would provide more comprehensive and precise information about the state of health of a test animal would be highly desirable. Such additional assays would preferably be less expensive, more rapid, and more convenient than current testing procedures, and would at the same time provide enough information to make early judgments regarding the safety of new compounds.

## Summary of the Invention

An object of the invention is to provide a new approach to toxicity assessment based on an examination of gene expression patterns, or profiles, in in vitro or in vivo test systems.

Another object of the invention is to provide a database on which to base decisions concerning the toxicological properties of chemicals, particularly drug candidates.

A further object of the invention is to provide a method for analyzing gene expression patterns in selected tissues of test animals.

A still further object of the invention is to provide a system for identifying genes which are differentially expressed in response to exposure to a test compound.

Another object of the invention is to provide a rapid and reliable method for correlating gene expression with short term and long term toxicity in test animals.

Another object of the invention is to identify genes whose expression is predictive of deleterious toxicity.

The invention achieves these and other objects by providing a method for massively parallel signature sequencing of genes expressed in one or more selected tissues of an organism exposed to a test compound. An important feature of the invention is the application of novel DNA sorting and sequencing methodologies that permit the formation of gene expression profiles for selected tissues by determining the sequence of portions of many thousands of different polynucleotides in parallel. Such profiles may be compared with those from tissues of control organisms at single or multiple time points to identify expression patterns predictive of toxicity.

The sorting methodology of the invention makes use of oligonucleotide tags that are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set cannot form a duplex (or triplex) with the complement of any other member with less than two mismatches. Complements of oligonucleotide tags of the invention, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements provide a means of enhancing specificity of hybridization for sorting polynucleotides, such as cDNAs.

The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully below, this condition is achieved by employing a repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions. The sorted populations of polynucleotides can then be sequenced on the solid phase support by a "single-base" or "base-by-base" sequencing methodology, as described more fully below.

In one aspect, the method of the invention comprises the following steps: (a) administering the compound to a test organism; (b) extracting a population of mRNA molecules from each of one or more tissues of the test organism; (c) forming a separate population of cDNA molecules from each population of mRNA molecules extracted from the one or more tissues such that each cDNA molecule of the separate populations has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set; (d) separately sampling each population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached; (e) sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports; (f) determining the nucleotide sequence of a portion of each of the sorted cDNA molecules of each separate population to form a frequency distribution of expressed genes for each of

the one or more tissues; and (g) correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

An important aspect of the invention is the identification of genes whose expression is predictive of the toxicity of a compound. Once such genes are identified, they may be employed in conventional assays, such as reverse transcriptase polymerase chain reaction (RT-PCR) assays for gene expression.

## Brief Description of the Drawings

Figure 1 is a flow chart representation of an algorithm for generating minimally cross-hybridizing sets of oligonucleotides.

Figure 2 diagrammatically illustrates an apparatus for carrying out polynucleotide sequencing in accordance with the invention.

## Definitions

"Complement" or "tag complement" as used herein in reference to oligonucleotide tags refers to an oligonucleotide to which a oligonucleotide tag specifically hybridizes to form a perfectly matched duplex or triplex. In embodiments where specific hybridization results in a triplex, the oligonucleotide tag may be selected to be either double stranded or single stranded. Thus, where triplexes are formed, the term "complement" is meant to encompass either a double stranded complement of a single stranded oligonucleotide tag or a single stranded complement of a double stranded oligonucleotide tag.

The term "oligonucleotide" as used herein includes linear oligomers of natural or modified monomers or linkages, including deoxyribonucleosides, ribonucleosides, anomeric forms thereof, peptide nucleic acids (PNAs), and the like, capable of specifically binding to a target polynucleotide by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like. Usually monomers are linked by phosphodiester bonds or analogs thereof to form oligonucleotides ranging in size from a few monomeric units, e.g. 3-4, to several tens of monomeric units. Whenever an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless otherwise noted. Analogs of phosphodiester linkages include phosphorothioate, phosphorodithioate, phosphoranilidate, phosphoramidate, and the like. Usually oligonucleotides of the invention comprise the four natural nucleotides; however, they may also comprise non-natural nucleotide analogs. It is clear to those skilled in the

art when oligonucleotides having natural or non-natural nucleotides may be employed, e.g. where processing by enzymes is called for, usually oligonucleotides consisting of natural nucleotides are required.

"Perfectly matched" in reference to a duplex means that the poly- or oligonucleotide strands making up the duplex form a double stranded structure with one other such that every nucleotide in each strand undergoes Watson-Crick basepairing with a nucleotide in the other strand. The term also comprehends the pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine bases, and the like, that may be employed. In reference to a triplex, the term means that the triplex consists of a perfectly matched duplex and a third strand in which every nucleotide undergoes Hoogsteen or reverse Hoogsteen association with a basepair of the perfectly matched duplex. Conversely, a "mismatch" in a duplex between a tag and an oligonucleotide means that a pair or triplet of nucleotides in the duplex or triplex fails to undergo Watson-Crick and/or Hoogsteen and/or reverse Hoogsteen bonding.

As used herein, "nucleoside" includes the natural nucleosides, including 2'-deoxy and 2'-hydroxyl forms, e.g. as described in Kornberg and Baker, DNA Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Analogs" in reference to nucleosides includes synthetic nucleosides having modified base moieties and/or modified sugar moieties, e.g. described by Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990), or the like, with the only proviso that they are capable of specific hybridization. Such analogs include synthetic nucleosides designed to enhance binding properties, reduce complexity, increase specificity, and the like.

As used herein "sequence determination" or "determining a nucleotide sequence" in reference to polynucleotides includes determination of partial as well as full sequence information of the polynucleotide. That is, the term includes sequence comparisons, fingerprinting, and like levels of information about a target polynucleotide, as well as the express identification and ordering of nucleosides, usually each nucleoside, in a target polynucleotide. The term also includes the determination of the identification, ordering, and locations of one, two, or three of the four types of nucleotides within a target polynucleotide. For example, in some embodiments sequence determination may be effected by identifying the ordering and locations of a single type of nucleotide, e.g. cytosines, within the target polynucleotide "CATCGC ..." so that its sequence is represented as a binary code, e.g. "100101 ... " for "C-(not C)-(not C)-C-(not C)-C ... " and the like.

As used herein, the term "complexity" in reference to a population of polynucleotides means the number of different species of molecule present in the population.

As used herein, the terms "gene expression profile," and "gene expression pattern" which is used equivalently, means a frequency distribution of sequences of portions of cDNA molecules sampled from a population of tag-cDNA conjugates. Generally, the portions of sequence are sufficiently long to uniquely identify the cDNA from which the portion arose. Preferably, the total number of sequences determined is at least 1000; more preferably, the total number of sequences determined in a gene expression profile is at least ten thousand.

As used herein, "test organism" means any in vitro or in vivo system which provides measureable responses to exposure to test compounds. Typically, test organisms may be mammalian cell cultures, particularly of specific tissues, such as hepatocytes, neurons, kidney cells, colony forming cells, or the like, or test organisms may be whole animals, such as rats, mice, hamsters, guinea pigs, dogs, cats, rabbits, pigs, monkeys, and the like.

## Detailed Description of the Invention

The invention provides a method for determining the toxicity of a compound by analyzing changes in the gene expression profiles in selected tissues of test organisms exposed to the compound. The invention also provides a method of identifying toxicity markers consisting of individual genes or a group of genes that is expressed acutely and which is correlated with prolonged or chronic toxicity, or suggests that the compound will have an undesirable cross reactivity. Gene expression profiles are generated by sequencing portions of cDNA molecules construction from mRNA extracted from tissues of test organisms exposed to the compound being tested. As used herein, the term "tissue" is employed with its usual medical or biological meaning, except that in reference to an in vitro test system, such as a cell culture, it simply means a sample from the culture. Gene expression profiles derived from test organisms are compared to gene expression profiles derived from control organisms to determine the genes which are differentially expressed in the test organism because of exposure to the compound being tested. In both cases, the sequence information of the gene expression profiles is obtained by massively parallel signature sequencing of cDNAs, which is implemented in steps (c) through (f) of the above method.

## Toxicity Assessment

Procedures for designing and conducting toxicity tests in in vitro and in vivo systems is well known, and is described in many texts on the subject, such as Loomis

et al. Loomis's Esstentials of Toxicology, 4th Ed. (Academic Press, New York, 1996); Echobichon, The Basics of Toxicity Testing (CRC Press, Boca Raton, 1992); Frazier, editor, In Vitro Toxicity Testing (Marcel Dekker, New York, 1992); and the like.

In toxicity testing, two groups of test organisms are usually employed: one group serves as a control and the other group receives the test compound in a single dose (for acute toxicity tests) or a regimen of doses (for prolonged or chronic toxicity tests). Since in most cases, the extraction of tissue as called for in the method of the invention requires sacrificing the test animal, both the control group and the group receiving compound must be large enough to permit removal of animals for sampling tissues, if it is desired to observe the dynamics of gene expression through the duration of an experiment.

In setting up a toxicity study, extensive guidance is provided in the literature for selecting the appropriate test organism for the compound being tested, route of administration, dose ranges, and the like. Water or physiological saline (0.9% NaCl in water) is the solute of choice for the test compound since these solvents permit administration by a variety of routes. When this is not possible because of solubility limitations, it is necessary to resort to the use of vegetable oils such as corn oil or even organic solvents, of which propylene glycol is commonly used. Whenever possible the use of suspension of emulsion should be avoided except for oral administration. Regardless of the route of administration, the volume required to administer a given dose is limited by the size of the animal that is used. It is desirable to keep the volume of each dose uniform within and between groups of animals. When rates or mice are used the volume administered by the oral route should not exceed 0.005 ml per gram of animal. Even when aqueous or physiological saline solutions are used for parenteral injection the volumes that are tolerated are limited, although such solutions are ordinarily thought of as being innocuous. The intravenous $LD_{50}$ of distilled water in the mouse is approximately 0.044 ml per gram and that of isotonic saline is 0.068 ml per gram of mouse.

When a compound is to be administered by inhalation, special techniques for generating test atmospheres are necessary. Dose estimation becomes very complicated. The methods usually involve aerosolization or nebulization of fluids containing the compound. If the agent to be tested is a fluid that has an appreciable vapor pressure, it may be administered by passing air through the solution under controlled temperature conditions. Under these conditions, dose is estimated from the volume of air inhaled per unit time, the temperature of the solution, and the vapor pressure of the agent involved. Gases are metered from reservoirs. When particles of a solution are to be administered, unless the particle size is less than about 2 μm the particles will not reach the terminal alveolar sacs in the lungs. A variety of

apparatuses and chambers are available to perform studies for detecting effects of irritant or other toxic endpoints when they are administered by inhalation. The preferred method of administering an agent to animals is via the oral route, either by intubation or by incorporating the agent in the feed.

Preferably, in designing a toxicity assessment, two or more species should be employed that handle the test compound as similarly to man as possible in terms of metabolism, absorption, excretion, tissue storage, and the like. Preferably, multiple doses or regimens at different concentrations should be employed to establish a dose-response relationship with respect to toxic effects. And preferably, the route of administration to the test animal should be the same as, or as similar as possible to, the route of administration of the compound to man. Effects obtained by one route of administration to test animals are not a priori applicable to effects by another route of administration to man. For example, food additives for man should be tested by admixture of the material in the diet of the test animals.

Acute toxicity tests consist of administering a compound to test organisms on one occasion. The purpose of such test is to determine the symptomotology consequent to administration of the compound and to determine the degree of lethality of the compound. The initial procedure is to perform a series of range-finding doses of the compound in a single species. This necessitates selection of a route of administration, preparation of the compound in a form suitable for administration by the selected route, and selection of an appropriate species. Preferably, initial acute toxicity studies are performed on either rats or mice because of their low cost, their availability, and the availability of abundant toxicologic reference data on these species. Prolonged toxicity tests consist of administering a compound to test organisms repeatedly, usually on a daily basis, over a period of 3 to 4 months. Two practical factors are encountered that place constraints on the design of such tests: First, the available routes of administration are limited because the route selected must be suitable for repeated administration without inducing harmful effects. And second, blood, urine, and perhaps other samples, should be taken repeatedly without inducing significant harm to the test animals. Preferably, in the method of the invention the gene expression profiles are obtained in conjunction with the measurement of the traditional toxicologic parameters, such as listed in the table below:

| Hematology | Blood Chemistry | Urine Analyses |
|---|---|---|
| erythrocyte count | sodium | pH |
| total leukocyte count | potassium | specific gravity |
| differential leukocyte count | chloride | total protein |
| hematocrit | calcium | sediment |
| hemoglobin | carbon dioxide | glucose |
| | serum glutamine-pyruvate transaminase | ketones |
| | serum glutamin-oxalacetic transaminase | bilirubin |
| | serum protein electrophoresis | |
| | blood sugar | |
| | blood urea nitrogen | |
| | total serum protein | |
| | serum albumin | |
| | total serum bilirubin | |

Oligonucleotide Tags and Tag Complements

Oligonucleotide tags are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set cannot form a duplex (or triplex) with the complement of any other member with less than two mismatches. Complements of oligonucleotide tags, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements provide a means of enhancing specificity of hybridization for sorting, tracking, or labeling molecules, especially polynucleotides.

Minimally cross-hybridizing sets of oligonucleotide tags and tag complements may be synthesized either combinatorially or individually depending on the size of the set desired and the degree to which cross-hybridization is sought to be minimized (or stated another way, the degree to which specificity is sought to be enhanced). For example, a minimally cross-hybridizing set may consist of a set of individually synthesized 10-mer sequences that differ from each other by at least 4 nucleotides, such set having a maximum size of 332 (when composed of 3 kinds of nucleotides and counted using a computer program such as disclosed in Appendix Ic). Alternatively, a minimally cross-hybridizing set of oligonucleotide tags may also be

assembled combinatorially from subunits which themselves are selected from a minimally cross-hybridizing set. For example, a set of minimally cross-hybridizing 12-mers differing from one another by at least three nucleotides may be synthesized by assembling 3 subunits selected from a set of minimally cross-hybridizing 4-mers that each differ from one another by three nucleotides. Such an embodiment gives a maximally sized set of $9^3$, or 729, 12-mers. The number 9 is number of oligonucleotides listed by the computer program of Appendix Ia, which assumes, as with the 10-mers, that only 3 of the 4 different types of nucleotides are used. The set is described as "maximal" because the computer programs of Appendices Ia-c provide the largest set for a given input (e.g. length, composition, difference in number of nucleotides between members). Additional minimally cross-hybridizing sets may be formed from subsets of such calculated sets.

Oligonucleotide tags may be single stranded and be designed for specific hybridization to single stranded tag complements by duplex formation or for specific hybridization to double stranded tag complements by triplex formation. Oligonucleotide tags may also be double stranded and be designed for specific hybridization to single stranded tag complements by triplex formation.

When synthesized combinatorially, an oligonucleotide tag preferably consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length wherein each subunit is selected from the same minimally cross-hybridizing set. In such embodiments, the number of oligonucleotide tags available depends on the number of subunits per tag and on the length of the subunits. The number is generally much less than the number of all possible sequences the length of the tag. which for a tag n nucleotides long would be $4^n$.

Complements of oligonucleotide tags attached to a solid phase support are used to sort polynucleotides from a mixture of polynucleotides each containing a tag. Complements of the oligonucleotide tags are synthesized on the surface of a solid phase support, such as a microscopic bead or a specific location on an array of synthesis locations on a single support, such that populations of identical sequences are produced in specific regions. That is, the surface of each support, in the case of a bead, or of each region, in the case of an array, is derivatized by only one type of complement which has a particular sequence. The population of such beads or regions contains a repertoire of complements with distinct sequences. As used herein in reference to oligonucleotide tags and tag complements, the term "repertoire" means the set of minimally cross-hybridizing set of oligonucleotides that make up the tags in a particular embodiment or the corresponding set of tag complements.

The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully

below, this condition is achieved by employing a repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions.

The nucleotide sequences of oligonucleotides of a minimally cross-hybridizing set are conveniently enumerated by simple computer programs, such as those exemplified by programs whose source codes are listed in Appendices Ia and Ib. Program minhx of Appendix Ia computes all minimally cross-hybridizing sets having 4-mer subunits composed of three kinds of nucleotides. Program tagN of Appendix Ib enumerates longer oligonucleotides of a minimally cross-hybridizing set. Similar algorithms and computer programs are readily written for listing oligonucleotides of minimally cross-hybridizing sets for any embodiment of the invention. Table I below provides guidance as to the size of sets of minimally cross-hybridizing oligonucleotides for the indicated lengths and number of nucleotide differences. The above computer programs were used to generate the numbers.

Table I

| Oligonucleotide Word Length | Nucleotide Difference between Oligonucleotides of Minimally Cross-Hybridizing Set | Maximal Size of Minimally Cross-Hybridizing Set | Size of Repertoire with Four Words | Size of Repertoire with Five Words |
|---|---|---|---|---|
| 4 | 3 | 9 | 6561 | $5.90 \times 10^4$ |
| 6 | 3 | 27 | $5.3 \times 10^5$ | $1.43 \times 10^7$ |
| 7 | 4 | 27 | $5.3 \times 10^5$ | $1.43 \times 10^7$ |
| 7 | 5 | 8 | 4096 | $3.28 \times 10^4$ |
| 8 | 3 | 190 | $1.30 \times 10^9$ | $2.48 \times 10^{11}$ |
| 8 | 4 | 62 | $1.48 \times 10^7$ | $9.16 \times 10^8$ |
| 8 | 5 | 18 | $1.05 \times 10^5$ | $1.89 \times 10^6$ |
| 9 | 5 | 39 | $2.31 \times 10^6$ | $9.02 \times 10^7$ |
| 10 | 5 | 332 | $1.21 \times 10^{10}$ | |
| 10 | 6 | 28 | $6.15 \times 10^5$ | $1.72 \times 10^7$ |
| 11 | 5 | 187 | | |
| 18 | 6 | ≈25000 | | |

18 12 24

For some embodiments of the invention, where extremely large repertoires of tags are not required, oligonucleotide tags of a minimally cross-hybridizing set may be separately synthesized. Sets containing several hundred to several thousands, or even several tens of thousands, of oligonucleotides may be synthesized directly by a variety of parallel synthesis approaches, e.g. as disclosed in Frank et al, U.S. patent 4,689,405; Frank et al, Nucleic Acids Research, 11: 4365-4377 (1983); Matson et al, Anal. Biochem., 224: 110-116 (1995); Fodor et al, International application PCT/US93/04145; Pease et al, Proc. Natl. Acad. Sci., 91: 5022-5026 (1994); Southern et al, J. Biotechnology, 35: 217-227 (1994), Brennan, International application PCT/US94/05896; Lashkari et al, Proc. Natl. Acad. Sci., 92: 7912-7915 (1995); or the like.

Preferably, oligonucleotide tags of the invention are synthesized combinatorially out of subunits between three and six nucleotides in length and selected from the same minimally cross-hybridizing set. For oligonucletides in this range, the members of such sets may be enumerated by computer programs based on the algorithm of Fig. 1.

The algorithm of Fig. 1 is implemented by first defining the characteristics of the subunits of the minimally cross-hybridizing set, i.e. length, number of base differences between members, and composition, e.g. do they consist of two, three, or four kinds of bases. A table $M_n$, n=1, is generated (100) that consists of all possible sequences of a given length and composition. An initial subunit $S_1$ is selected and compared (120) with successive subunits $S_i$ for i=n+1 to the end of the table. Whenever a successive subunit has the required number of mismatches to be a member of the minimally cross-hybridizing set, it is saved in a new table $M_{n+1}$ (125), that also contains subunits previously selected in prior passes through step 120. For example, in the first set of comparisons, $M_2$ will contain $S_1$; in the second set of comparisons, $M_3$ will contain $S_1$ and $S_2$; in the third set of comparisons, $M_4$ will contain $S_1$, $S_2$, and $S_3$; and so on. Similarly, comparisons in table $M_j$ will be between $S_j$ and all successive subunits in $M_j$. Note that each successive table $M_{n+1}$ is smaller than its predecessors as subunits are eliminated in successive passes through step 130. After every subunit of table $M_n$ has been compared (140) the old table is replaced by the new table $M_{n+1}$, and the next round of comparisons are begun. The process stops (160) when a table $M_n$ is reached that contains no successive subunits to compare to the selected subunit $S_i$, i.e. $M_n=M_{n+1}$.

Preferably, minimally cross-hybridizing sets comprise subunits that make approximately equivalent contributions to duplex stability as every other subunit in

the set. In this way, the stability of perfectly matched duplexes between every subunit and its complement is approximately equal. Guidance for selecting such sets is provided by published techniques for selecting optimal PCR primers and calculating duplex stabilities, e.g. Rychlik et al, Nucleic Acids Research, 17: 8543-8551 (1989) and 18: 6409-6412 (1990); Breslauer et al, Proc. Natl. Acad. Sci., 83: 3746-3750 (1986); Wetmur, Crit. Rev. Biochem. Mol. Biol., 26: 227-259 (1991);and the like. For shorter tags, e.g. about 30 nucleotides or less, the algorithm described by Rychlik and Wetmur is preferred, and for longer tags, e.g. about 30-35 nucleotides or greater, an algorithm disclosed by Suggs et al, pages 683-693 in Brown, editor, ICN-UCLA Symp. Dev. Biol., Vol. 23 (Academic Press, New York, 1981) may be conveniently employed. Clearly, the are many approaches available to one skilled in the art for designing sets of minimally cross-hybridizing subunits within the scope of the invention. For example, to minimize the affects of different base-stacking energies of terminal nucleotides when subunits are assembled, subunits may be provided that have the same terminal nucleotides. In this way, when subunits are linked, the sum of the base-stacking energies of all the adjoining terminal nucleotides will be the same, thereby reducing or eliminating variability in tag melting temperatures.

A "word" of terminal nucleotides, shown in italic below, may also be added to each end of a tag so that a perfect match is always formed between it and a similar terminal "word" on any other tag complement. Such an augmented tag would have the form:

| | | | | | | |
|---|---|---|---|---|---|---|
| *W* | $W_1$ | $W_2$ | ... | $W_{k-1}$ | $W_k$ | *W* |
| *W'* | $W_1$' | $W_2$' | ... | $W_{k-1}$' | $W_k$' | *W'* |

where the primed W's indicate complements. With ends of tags always forming perfectly matched duplexes, all mismatched words will be internal mismatches thereby reducing the stability of tag-complement duplexes that otherwise would have mismatched words at their ends. It is well known that duplexes with internal mismatches are significantly less stable than duplexes with the same mismatch at a terminus.

A preferred embodiment of minimally cross-hybridizing sets are those whose subunits are made up of three of the four natural nucleotides. As will be discussed more fully below, the absence of one type of nucleotide in the oligonucleotide tags permits target polynucleotides to be loaded onto solid phase supports by use of the 5'→3' exonuclease activity of a DNA polymerase. The following is an exemplary minimally cross-hybridizing set of subunits each comprising four nucleotides selected from the group consisting of A, G, and T:

Table II

| Word: | $w_1$ | $w_2$ | $w_3$ | $w_4$ |
|---|---|---|---|---|
| Sequence: | GATT | TGAT | TAGA | TTTG |
| Word: | $w_5$ | $w_6$ | $w_7$ | $w_8$ |
| Sequence: | GTAA | AGTA | ATGT | AAAG |

In this set, each member would form a duplex having three mismatched bases with the complement of every other member.

Further exemplary minimally cross-hybridizing sets are listed below in Table III. Clearly, additional sets can be generated by substituting different groups of nucleotides, or by using subsets of known minimally cross-hybridizing sets.

Table III

Exemplary Minimally Cross-Hybridizing Sets of 4-mer Subunits

| Set 1 | Set 2 | Set 3 | Set 4 | Set 5 | Set 6 |
|---|---|---|---|---|---|
| CATT | ACCC | AAAC | AAAG | AACA | AACG |
| CTAA | AGGG | ACCA | ACCA | ACAC | ACAA |
| TCAT | CACG | AGGG | AGGC | AGGG | AGGC |
| ACTA | CCGA | CACG | CACC | CAAG | CAAC |
| TACA | CGAC | CCGC | CCGG | CCGC | CCGG |
| TTTC | GAGC | CGAA | CGAA | CGCA | CGCA |
| ATCT | GCAG | GAGA | GAGA | GAGA | GAGA |
| AAAC | GGCA | GCAG | GCAC | GCCG | GCCC |
| | AAAA | GGCC | GGCG | GGAC | GGAG |

| Set 7 | Set 8 | Set 9 | Set 10 | Set 11 | Set 12 |
|---|---|---|---|---|---|
| AAGA | AAGC | AAGG | ACAG | ACCG | ACGA |
| ACAC | ACAA | ACAA | AACA | AAAA | AAAC |
| AGCG | AGCG | AGCC | AGGC | AGGC | AGCG |
| CAAG | CAAG | CAAC | CAAC | CACC | CACA |
| CCCA | CCCC | CCCG | CCGA | CCGA | CCAG |
| CGGC | CGGA | CGGA | CGCG | CGAG | CGGC |
| GACC | GACA | GACA | GAGG | GAGG | GAGG |
| GCGG | GCGG | GCGC | GCCC | GCAC | GCCC |
| GGAA | GGAC | GGAG | GGAA | GGCA | GGAA |

The oligonucleotide tags of the invention and their complements are conveniently synthesized on an automated DNA synthesizer, e.g. an Applied Biosystems, Inc. (Foster City, California) model 392 or 394 DNA/RNA Synthesizer, using standard chemistries, such as phosphoramidite chemistry, e.g. disclosed in the following references: Beaucage and Iyer, Tetrahedron, 48: 2223-2311 (1992); Molko et al, U.S. patent 4,980,460; Koster et al, U.S. patent 4,725,677; Caruthers et al, U.S. patents 4,415,732; 4,458,066; and 4,973,679; and the like. Alternative chemistries, e.g. resulting in non-natural backbone groups, such as phosphorothioate, phosphoramidate, and the like, may also be employed provided that the resulting oligonucleotides are capable of specific hybridization. In some embodiments, tags may comprise naturally occurring nucleotides that permit processing or manipulation by enzymes, while the corresponding tag complements may comprise non-natural nucleotide analogs, such as peptide nucleic acids, or like compounds, that promote the formation of more stable duplexes during sorting.

When microparticles are used as supports, repertoires of oligonucleotide tags and tag complements may be generated by subunit-wise synthesis via "split and mix" techniques, e.g. as disclosed in Shortle et al, International patent application PCT/US93/03418 or Lyttle et al, Biotechniques, 19: 274-280 (1995). Briefly, the basic unit of the synthesis is a subunit of the oligonucleotide tag. Preferably, phosphoramidite chemistry is used and 3' phosphoramidite oligonucleotides are prepared for each subunit in a minimally cross-hybridizing set, e.g. for the set first listed above, there would be eight 4-mer 3'-phosphoramidites. Synthesis proceeds as disclosed by Shortle et al or in direct analogy with the techniques employed to generate diverse oligonucleotide libraries using nucleosidic monomers, e.g. as disclosed in Telenius et al, Genomics, 13: 718-725 (1992); Welsh et al, Nucleic Acids Research, 19: 5275-5279 (1991); Grothues et al, Nucleic Acids Research, 21: 1321-1322 (1993); Hartley, European patent application 90304496.4; Lam et al, Nature, 354: 82-84 (1991); Zuckerman et al, Int. J. Pept. Protein Research, 40: 498-507 (1992); and the like. Generally, these techniques simply call for the application of

mixtures of the activated monomers to the growing oligonucleotide during the coupling steps. Preferably, oligonucleotide tags and tag complements are synthesized on a DNA synthesizer having a number of synthesis chambers which is greater than or equal to the number of different kinds of words used in the construction of the tags. That is, preferably there is a synthesis chamber corresponding to each type of word. In this embodiment, words are added nucleotide-by-nucleotide, such that if a word consists of five nucleotides there are five monomer couplings in each synthesis chamber. After a word is completely synthesized, the synthesis supports are removed from the chambers, mixed, and redistributed back to the chambers for the next cycle of word addition. This latter embodiment takes advantage of the high coupling yields of monomer addition, e.g. in phosphoramidite chemistries.

Double stranded forms of tags may be made by separately synthesizing the complementary strands followed by mixing under conditions that permit duplex formation. Alternatively, double stranded tags may be formed by first synthesizing a single stranded repertoire linked to a known oligonucleotide sequence that serves as a primer binding site. The second strand is then synthesized by combining the single stranded repertoire with a primer and extending with a polymerase. This latter approach is described in Oliphant et al, Gene, 44: 177-183 (1986). Such duplex tags may then be inserted into cloning vectors along with target polynucleotides for sorting and manipulation of the target polynucleotide in accordance with the invention.

When tag complements are employed that are made up of nucleotides that have enhanced binding characteristics, such as PNAs or oligonucleotide N3'→P5' phosphoramidates, sorting can be implemented through the formation of D-loops between tags comprising natural nucleotides and their PNA or phosphoramidate complements, as an alternative to the "stripping" reaction employing the 3'→5' exonuclease activity of a DNA polymerase to render a tag single stranded.

Oligonucleotide tags of the invention may range in length from 12 to 60 nucleotides or basepairs. Preferably, oligonucleotide tags range in length from 18 to 40 nucleotides or basepairs. More preferably, oligonucleotide tags range in length from 25 to 40 nucleotides or basepairs. In terms of preferred and more preferred numbers of subunits, these ranges may be expressed as follows:

Table IV

Numbers of Subunits in Tags in Preferred Embodiments

| Monomers in Subunit | Nucleotides in Oligonucleotide Tag | | |
|---|---|---|---|
| | (12-60) | (18-40) | (25-40) |

| | | | |
|---|---|---|---|
| 3 | 4-20 subunits | 6-13 subunits | 8-13 subunits |
| 4 | 3-15 subunits | 4-10 subunits | 6-10 subunits |
| 5 | 2-12 subunits | 3-8 subunits | 5-8 subunits |
| 6 | 2-10 subunits | 3-6 subunits | 4-6 subunits |

Most preferably, oligonucleotide tags are single stranded and specific hybridization occurs via Watson-Crick pairing with a tag complement.

Preferably, repertoires of single stranded oligonucleotide tags of the invention contain at least 100 members; more preferably, repertoires of such tags contain at least 1000 members; and most preferably, repertoires of such tags contain at least 10,000 members.

Triplex Tags

In embodiments where specific hybridization occurs via triplex formation, coding of tag sequences follows the same principles as for duplex-forming tags; however, there are further constraints on the selection of subunit sequences. Generally, third strand association via Hoogsteen type of binding is most stable along homopyrimidine-homopurine tracks in a double stranded target. Usually, base triplets form in T-A*T or C-G*C motifs (where "-" indicates Watson-Crick pairing and "*" indicates Hoogsteen type of binding); however, other motifs are also possible. For example, Hoogsteen base pairing permits parallel and antiparallel orientations between the third strand (the Hoogsteen strand) and the purine-rich strand of the duplex to which the third strand binds, depending on conditions and the composition of the strands. There is extensive guidance in the literature for selecting appropriate sequences, orientation, conditions, nucleoside type (e.g. whether ribose or deoxyribose nucleosides are employed), base modifications (e.g. methylated cytosine, and the like) in order to maximize, or otherwise regulate, triplex stability as desired in particular embodiments, e.g. Roberts et al, Proc. Natl. Acad. Sci., 88: 9397-9401 (1991); Roberts et al, Science, 258: 1463-1466 (1992); Roberts et al, Proc. Natl. Acad. Sci., 93: 4320-4325 (1996); Distefano et al, Proc. Natl. Acad. Sci., 90: 1179-1183 (1993); Mergny et al, Biochemistry, 30: 9791-9798 (1991); Cheng et al, J. Am. Chem. Soc., 114: 4465-4474 (1992); Beal and Dervan, Nucleic Acids Research, 20: 2773-2776 (1992); Beal and Dervan, J. Am. Chem. Soc., 114: 4976-4982 (1992); Giovannangeli et al, Proc. Natl. Acad. Sci., 89: 8631-8635 (1992); Moser and Dervan, Science, 238: 645-650 (1987); McShan et al, J. Biol. Chem., 267:5712-5721 (1992); Yoon et al, Proc. Natl. Acad. Sci., 89: 3840-3844 (1992); Blume et al, Nucleic Acids Research, 20: 1777-1784 (1992); Thuong and Helene, Angew. Chem. Int. Ed. Engl.

32: 666-690 (1993); Escude et al, Proc. Natl. Acad. Sci., 93: 4365-4369 (1996); and the like. Conditions for annealing single-stranded or duplex tags to their single-stranded or duplex complements are well known, e.g. Ji et al, Anal. Chem. 65: 1323-1328 (1993); Cantor et al, U.S. patent 5,482,836; and the like. Use of triplex tags has the advantage of not requiring a "stripping" reaction with polymerase to expose the tag for annealing to its complement.

Preferably, oligonucleotide tags of the invention employing triplex hybridization are double stranded DNA and the corresponding tag complements are single stranded. More preferably, 5-methylcytosine is used in place of cytosine in the tag complements in order to broaden the range of pH stability of the triplex formed between a tag and its complement. Preferred conditions for forming triplexes are fully disclosed in the above references. Briefly, hybridization takes place in concentrated salt solution, e.g. 1.0 M NaCl, 1.0 M potassium acetate, or the like, at pH below 5.5 ( or 6.5 if 5-methylcytosine is employed). Hybridization temperature depends on the length and composition of the tag; however, for an 18-20-mer tag of longer, hybridization at room temperature is adequate. Washes may be conducted with less concentrated salt solutions, e.g. 10 mM sodium acetate, 100 mM $MgCl_2$, pH 5.8, at room temperature. Tags may be eluted from their tag complements by incubation in a similar salt solution at pH 9.0.

Minimally cross-hybridizing sets of oligonucleotide tags that form triplexes may be generated by the computer program of Appendix Ic, or similar programs. An exemplary set of double stranded 8-mer words are listed below in capital letters with the corresponding complements in small letters. Each such word differs from each of the other words in the set by three base pairs.

Table V

Exemplary Minimally Cross-Hybridizing Set of DoubleStranded 8-mer Tags

| | | | |
|---|---|---|---|
| 5'-AAGGAGAG<br>3'-TTCCTCTC<br>3'-ttcctctc | 5'-AAAGGGGA<br>3'-TTTCCCCT<br>3'-tttcccct | 5'-AGAGAAGA<br>3'-TCTCTTCT<br>3'-tctcttct | 5'-AGGGGGGG<br>3'-TCCCCCCC<br>3'-tccccccc |
| 5'-AAAAAAAA<br>3'-TTTTTTTT<br>3'-tttttttt | 5'-AAGAGAGA<br>3'-TTCTCTCT<br>3'-ttctctct | 5'-AGGAAAAG<br>3'-TCCTTTTC<br>3'-tcctttte | 5'-GAAAGGAG<br>3'-CTTTCCTC<br>3'-ctttcctc |
| 5'-AAAAAGGG<br>3'-TTTTTCCC<br>3'-ttttccc | 5'-AGAAGAGG<br>3'-TCTTCTCC<br>3'-tcttctcc | 5'-AGGAAGGA<br>3'-TCCTTCCT<br>3'-tccttcct | 5'-GAAGAAGG<br>3'-CTTCTTCC<br>3'-cttcttcc |
| 5'-AAAGGAAG<br>3'-TTTCCTTC<br>3'-tttccttc | 5'-AGAAGGAA<br>3'-TCTTCCTT<br>3'-tcttcctt | 5'-AGGGGAAA<br>3'-TCCCCTTT<br>3'-tccccttt | 5'-GAAGAGAA<br>3'-CTTCTCTT<br>3'-cttctctt |

Table VI
Repertoire Size of Various Double Stranded Tags
That Form Triplexes with Their Tag Complements

| Oligonucleotide Word Length | Nucleotide Difference between Oligonucleotides of Minimally Cross-Hybridizing Set | Maximal Size of Minimally Cross-Hybridizing Set | Size of Repertoire with Four Words | Size of Repertoire with Five Words |
|---|---|---|---|---|
| 4 | 2 | 8 | 4096 | $3.2 \times 10^4$ |
| 6 | 3 | 8 | 4096 | $3.2 \times 10^4$ |
| 8 | 3 | 16 | $6.5 \times 10^4$ | $1.05 \times 10^6$ |
| 10 | 5 | 8 | 4096 | |
| 15 | 5 | 92 | | |
| 20 | 6 | 765 | | |
| 20 | 8 | 92 | | |
| 20 | 10 | 22 | | |

Preferably, repertoires of double stranded oligonucleotide tags of the invention contain at least 10 members; more preferably, repertoires of such tags contain at least 100 members. Preferably, words are between 4 and 8 nucleotides in length for combinatorially synthesized double stranded oligonucletide tags, and oligonucleotide tags are between 12 and 60 base pairs in length. More preferably, such tags are between 18 and 40 base pairs in length.

## Solid Phase Supports

Solid phase supports for use with the invention may have a wide variety of forms, including microparticles, beads, and membranes, slides, plates, micromachined chips, and the like. Likewise, solid phase supports of the invention may comprise a

wide variety of compositions, including glass, plastic, silicon, alkanethiolate-derivatized gold, cellulose, low cross-linked and high cross-linked polystyrene, silica gel, polyamide, and the like. Preferably, either a population of discrete particles are employed such that each has a uniform coating, or population, of complementary sequences of the same tag (and no other), or a single or a few supports are employed with spatially discrete regions each containing a uniform coating, or population, of complementary sequences to the same tag (and no other). In the latter embodiment, the area of the regions may vary according to particular applications; usually, the regions range in area from several $\mu m^2$, e.g. 3-5, to several hundred $\mu m^2$, e.g. 100-500. Preferably, such regions are spatially discrete so that signals generated by events, e.g. fluorescent emissions, at adjacent regions can be resolved by the detection system being employed. In some applications, it may be desirable to have regions with uniform coatings of more than one tag complement, e.g. for simultaneous sequence analysis, or for bringing separately tagged molecules into close proximity.

Tag complements may be used with the solid phase support that they are synthesized on, or they may be separately synthesized and attached to a solid phase support for use, e.g. as disclosed by Lund et al, Nucleic Acids Research, 16: 10861-10880 (1988); Albretsen et al, Anal. Biochem., 189: 40-50 (1990); Wolf et al, Nucleic Acids Research, 15: 2911-2926 (1987); or Ghosh et al, Nucleic Acids Research, 15: 5353-5372 (1987). Preferably, tag complements are synthesized on and used with the same solid phase support, which may comprise a variety of forms and include a variety of linking moieties. Such supports may comprise microparticles or arrays, or matrices, of regions where uniform populations of tag complements are synthesized. A wide variety of microparticle supports may be used with the invention, including microparticles made of controlled pore glass (CPG), highly cross-linked polystyrene, acrylic copolymers, cellulose, nylon, dextran, latex, polyacrolein, and the like, disclosed in the following exemplary references: Meth. Enzymol., Section A, pages 11-147, vol. 44 (Academic Press, New York, 1976); U.S. patents 4,678,814; 4,413,070; and 4,046;720; and Pon, Chapter 19, in Agrawal, editor, Methods in Molecular Biology, Vol. 20, (Humana Press, Totowa, NJ, 1993). Microparticle supports further include commercially available nucleoside-derivatized CPG and polystyrene beads (e.g. available from Applied Biosystems, Foster City, CA); derivatized magnetic beads; polystyrene grafted with polyethylene glycol (e.g., TentaGel™, Rapp Polymere, Tubingen Germany); and the like. Selection of the support characteristics, such as material, porosity, size, shape, and the like, and the type of linking moiety employed depends on the conditions under which the tags are used. For example, in applications involving successive processing with enzymes, supports and linkers that minimize steric hindrance of the enzymes and that facilitate

access to substrate are preferred. Other important factors to be considered in selecting the most appropriate microparticle support include size uniformity, efficiency as a synthesis support, degree to which surface area known, and optical properties, e.g. as explain more fully below, clear smooth beads provide instrumentational advantages when handling large numbers of beads on a surface.

Exemplary linking moieties for attaching and/or synthesizing tags on microparticle surfaces are disclosed in Pon et al, Biotechniques, 6:768-775 (1988); Webb, U.S. patent 4,659,774; Barany et al, International patent application PCT/US91/06103; Brown et al, J. Chem. Soc. Commun., 1989: 891-893; Damha et al, Nucleic Acids Research, 18: 3813-3821 (1990); Beattie et al, Clinical Chemistry, 39: 719-722 (1993); Maskos and Southern, Nucleic Acids Research, 20: 1679-1684 (1992); and the like.

As mentioned above, tag complements may also be synthesized on a single (or a few) solid phase support to form an array of regions uniformly coated with tag complements. That is, within each region in such an array the same tag complement is synthesized. Techniques for synthesizing such arrays are disclosed in McGall et al, International application PCT/US93/03767; Pease et al, Proc. Natl. Acad. Sci., 91: 5022-5026 (1994); Southern and Maskos, International application PCT/GB89/01114; Maskos and Southern (cited above); Southern et al, Genomics, 13: 1008-1017 (1992); and Maskos and Southern, Nucleic Acids Research, 21: 4663-4669 (1993).

Preferably, the invention is implemented with microparticles or beads uniformly coated with complements of the same tag sequence. Microparticle supports and methods of covalently or noncovalently linking oligonucleotides to their surfaces are well known, as exemplified by the following references: Beaucage and Iyer (cited above); Gait, editor, Oligonucleotide Synthesis: A Practical Approach (IRL Press, Oxford, 1984); and the references cited above. Generally, the size and shape of a microparticle is not critical; however, microparticles in the size range of a few, e.g. 1-2, to several hundred, e.g. 200-1000 μm diameter are preferable, as they facilitate the construction and manipulation of large repertoires of oligonucleotide tags with minimal reagent and sample usage.

In some preferred applications, commercially available controlled-pore glass (CPG) or polystyrene supports are employed as solid phase supports in the invention. Such supports come available with base-labile linkers and initial nucleosides attached, e.g. Applied Biosystems (Foster City, CA). Preferably, microparticles having pore size between 500 and 1000 angstroms are employed.

In other preferred applications, non-porous microparticles are employed for their optical properties, which may be advantageously used when tracking large

numbers of microparticles on planar supports, such as a microscope slide. Particularly preferred non-porous microparticles are the glycidal methacrylate (GMA) beads available from Bangs Laboratories (Carmel, IN). Such microparticles are useful in a variety of sizes and derivatized with a variety of linkage groups for synthesizing tags or tag complements. Preferably, for massively parallel manipulations of tagged microparticles, 5 μm diameter GMA beads are employed.

Attaching Tags to Polynucleotides
For Sorting onto Solid Phase Supports

An important aspect of the invention is the sorting and attachment of a populations of polynucleotides, e.g. from a cDNA library, to microparticles or to separate regions on a solid phase support such that each microparticle or region has substantially only one kind of polynucleotide attached. This objective is accomplished by insuring that substantially all different polynucleotides have different tags attached. This condition, in turn, is brought about by taking a sample of the full ensemble of tag-polynucleotide conjugates for analysis. (It is acceptable that identical polynucleotides have different tags, as it merely results in the same polynucleotide being operated on or analyzed twice in two different locations.) Such sampling can be carried out either overtly--for example, by taking a small volume from a larger mixture--after the tags have been attached to the polynucleotides, it can be carried out inherently as a secondary effect of the techniques used to process the polynucleotides and tags, or sampling can be carried out both overtly and as an inherent part of processing steps.

Preferably, in constructing a cDNA library where substantially all different cDNAs have different tags, a tag repertoire is employed whose complexity, or number of distinct tags, greatly exceeds the total number of mRNAs extracted from a cell or tissue sample. Preferably, the complexity of the tag repertoire is at least 10 times that of the polynucleotide population; and more preferably, the complexity of the tag repertoire is at least 100 times that of the polynucleotide population. Below, a protocol is disclosed for cDNA library construction using a primer mixture that contains a full repertoire of exemplary 9-word tags. Such a mixture of tag-containing primers has a complexity of $8^9$, or about $1.34 \times 10^8$. As indicated by Winslow et al, Nucleic Acids Research, 19: 3251-3253 (1991), mRNA for library construction can be extracted from as few as 10-100 mammalian cells. Since a single mammalian cell contains about $5 \times 10^5$ copies of mRNA molecules of about $3.4 \times 10^4$ different kinds,

by standard techniques one can isolate the mRNA from about 100 cells, or (theoretically) about $5 \times 10^7$ mRNA molecules. Comparing this number to the complexity of the primer mixture shows that without any additional steps, and even assuming that mRNAs are converted into cDNAs with perfect efficiency (1% efficiency or less is more accurate), the cDNA library construction protocol results in a population containing no more than 37% of the total number of different tags. That is, without any overt sampling step at all, the protocol inherently generates a sample that comprises 37%, or less, of the tag repertoire. The probability of obtaining a double under these conditions is about 5%, which is within the preferred range. With mRNA from 10 cells, the fraction of the tag repertoire sampled is reduced to only 3.7%, even assuming that all the processing steps take place at 100% efficiency. In fact, the efficiencies of the processing steps for constructing cDNA libraries are very low, a "rule of thumb" being that good library should contain about $10^8$ cDNA clones from mRNA extracted from $10^6$ mammalian cells.

Use of larger amounts of mRNA in the above protocol, or for larger amounts of polynucleotides in general, where the number of such molecules exceeds the complexity of the tag repertoire, a tag-polynucleotide conjugate mixture potentially contains every possible pairing of tags and types of mRNA or polynucleotide. In such cases, overt sampling may be implemented by removing a sample volume after a serial dilution of the starting mixture of tag-polynucleotide conjugates. The amount of dilution required depends on the amount of starting material and the efficiencies of the processing steps, which are readily estimated.

If mRNA were extracted from $10^6$ cells (which would correspond to about 0.5 μg of poly(A)$^+$ RNA), and if primers were present in about 10-100 fold concentration excess--as is called for in a typical protocol, e.g. Sambrook et al, Molecular Cloning, Second Edition, page 8.61 [10 μL 1.8 kb mRNA at 1 mg/mL equals about $1.68 \times 10^{-11}$ moles and 10 μL 18-mer primer at 1 mg/mL equals about $1.68 \times 10^{-9}$ moles], then the total number of tag-polynucleotide conjugates in a cDNA library would simply be equal to or less than the starting number of mRNAs, or about $5 \times 10^{11}$ vectors containing tag-polynucleotide conjugates--again this assumes that each step in cDNA construction--first strand synthesis, second strand synthesis, ligation into a vector--occurs with perfect efficiency, which is a very conservative estimate. The actual number is significantly less.

If a sample of n tag-polynucleotide conjugates are randomly drawn from a reaction mixture--as could be effected by taking a sample volume, the probability of drawing conjugates having the same tag is described by the Poisson distribution, $P(r)=e^{-\lambda}(\lambda)^r/r$, where r is the number of conjugates having the same tag and $\lambda=np$, where p is the probability of a given tag being selected. If $n=10^6$ and $p=1/(1.34 \times$

$10^8$), then $\lambda$=.00746 and P(2)=2.76 x $10^{-5}$. Thus, a sample of one million molecules gives rise to an expected number of doubles well within the preferred range. Such a sample is readily obtained as follows: Assume that the 5 x $10^{11}$ mRNAs are perfectly converted into 5 x $10^{11}$ vectors with tag-cDNA conjugates as inserts and that the 5 x $10^{11}$ vectors are in a reaction solution having a volume of 100 µl. Four 10-fold serial dilutions may be carried out by transferring 10 µl from the original solution into a vessel containing 90 µl of an appropriate buffer, such as TE. This process may be repeated for three additional dilutions to obtain a 100 µl solution containing 5 x $10^5$ vector molecules per µl. A 2 µl aliquot from this solution yields $10^6$ vectors containing tag-cDNA conjugates as inserts. This sample is then amplified by straight forward transformation of a competent host cell followed by culturing.

Of course, as mentioned above, no step in the above process proceeds with perfect efficiency. In particular, when vectors are employed to amplify a sample of tag-polynucleotide conjugates, the step of transforming a host is very inefficient. Usually, no more than 1% of the vectors are taken up by the host and replicated. Thus, for such a method of amplification, even fewer dilutions would be required to obtain a sample of $10^6$ conjugates.

A repertoire of oligonucleotide tags can be conjugated to a population of polynucleotides in a number of ways, including direct enzymatic ligation, amplification, e.g. via PCR, using primers containing the tag sequences, and the like. The initial ligating step produces a very large population of tag-polynucleotide conjugates such that a single tag is generally attached to many different polynucleotides. However, as noted above, by taking a sufficiently small sample of the conjugates, the probability of obtaining "doubles," i.e. the same tag on two different polynucleotides, can be made negligible. Generally, the larger the sample the greater the probability of obtaining a double. Thus, a design trade-off exists between selecting a large sample of tag-polynucleotide conjugates--which, for example, ensures adequate coverage of a target polynucleotide in a shotgun sequencing operation or adequate representation of a rapidly changing mRNA pool, and selecting a small sample which ensures that a minimal number of doubles will be present. In most embodiments, the presence of doubles merely adds an additional source of noise or, in the case of sequencing, a minor complication in scanning and signal processing, as microparticles giving multiple fluorescent signals can simply be ignored.

As used herein, the term "substantially all" in reference to attaching tags to molecules, especially polynucleotides, is meant to reflect the statistical nature of the sampling procedure employed to obtain a population of tag-molecule conjugates essentially free of doubles. The meaning of substantially all in terms of actual

percentages of tag-molecule conjugates depends on how the tags are being employed. Preferably, for nucleic acid sequencing, substantially all means that at least eighty percent of the polynucleotides have unique tags attached. More preferably, it means that at least ninety percent of the polynucleotides have unique tags attached. Still more preferably, it means that at least ninety-five percent of the polynucleotides have unique tags attached. And, most preferably, it means that at least ninety-nine percent of the polynucleotides have unique tags attached.

Preferably, when the population of polynucleotides consists of messenger RNA (mRNA), oligonucleotides tags may be attached by reverse transcribing the mRNA with a set of primers preferably containing complements of tag sequences. An exemplary set of such primers could have the following sequence (SEQ ID NO: 1):

5'-mRNA- $[A]_n$ -3'

$[T]_{19}$GG[W,W,W,C]$_9$ACCAGCTGATC-5'-biotin

where "[W,W,W,C]$_9$" represents the sequence of an oligonucleotide tag of nine subunits of four nucleotides each and "[W,W,W,C]" represents the subunit sequences listed above, i.e. "W" represents T or A. The underlined sequences identify an optional restriction endonuclease site that can be used to release the polynucleotide from attachment to a solid phase support via the biotin, if one is employed. For the above primer, the complement attached to a microparticle could have the form:

5'-[G,W,W,W]$_9$TGG-linker-microparticle

After reverse transcription, the mRNA is removed, e.g. by RNase H digestion, and the second strand of the cDNA is synthesized using, for example, a primer of the following form (SEQ ID NO: 2):

5'-NRRGATCYNNN-3'

where N is any one of A, T, G, or C; R is a purine-containing nucleotide, and Y is a pyrimidine-containing nucleotide. This particular primer creates a Bst Y1 restriction site in the resulting double stranded DNA which, together with the Sal I site, facilitates cloning into a vector with, for example, Bam HI and Xho I sites. After Bst Y1 and Sal I digestion, the exemplary conjugate would have the form:

5'-RCGACCA[C,W,W,W]$_9$GG[T]$_{19}$- cDNA -NNNR
GGT[G,W,W,W]$_9$CC[A]$_{19}$- rDNA -NNNYCTAG-5'

The polynucleotide-tag conjugates may then be manipulated using standard molecular biology techniques. For example, the above conjugate--which is actually a mixture--may be inserted into commercially available cloning vectors, e.g. Stratagene Cloning System (La Jolla, CA); transfected into a host, such as a commercially available host bacteria; which is then cultured to increase the number of conjugates. The cloning vectors may then be isolated using standard techniques, e.g. Sambrook et al, Molecular Cloning, Second Edition (Cold Spring Harbor Laboratory, New York, 1989). Alternatively, appropriate adaptors and primers may be employed so that the conjugate population can be increased by PCR.

Preferably, when the ligase-based method of sequencing is employed, the Bst Y1 and Sal I digested fragments are cloned into a Bam HI-/Xho I-digested vector having the following single-copy restriction sites (SEQ ID NO: 3):

5'-GAGGATGCCTTTATGGATCCACTCGAGATCCCAATCCA-3'
FokI BamHI XhoI

This adds the Fok I site which will allow initiation of the sequencing process discussed more fully below.

Tags can be conjugated to cDNAs of existing libraries by standard cloning methods. cDNAs are excised from their existing vector, isolated, and then ligated into a vector containing a repertoire of tags. Preferably, the tag-containing vector is linearized by cleaving with two restriction enzymes so that the excised cDNAs can be ligated in a predetermined orientation. The concentration of the linearized tag-containing vector is in substantial excess over that of the cDNA inserts so that ligation provides an inherent sampling of tags.

A general method for exposing the single stranded tag after amplification involves digesting a target polynucleotide-containing conjugate with the 5'→3' exonuclease activity of T4 DNA polymerase, or a like enzyme. When used in the presence of a single deoxynucleoside triphosphate, such a polymerase will cleave nucleotides from 3' recessed ends present on the non-template strand of a double stranded fragment until a complement of the single deoxynucleoside triphosphate is reached on the template strand. When such a nucleotide is reached the 5'→3' digestion effectively ceases, as the polymerase's extension activity adds nucleotides at a higher rate than the excision activity removes nucleotides. Consequently, single

stranded tags constructed with three nucleotides are readily prepared for loading onto solid phase supports.

The technique may also be used to preferentially methylate interior Fok I sites of a target polynucleotide while leaving a single Fok I site at the terminus of the polynucleotide unmethylated. First, the terminal Fok I site is rendered single stranded using a polymerase with deoxycytidine triphosphate. The double stranded portion of the fragment is then methylated, after which the single stranded terminus is filled in with a DNA polymerase in the presence of all four nucleoside triphosphates, thereby regenerating the Fok I site. Clearly, this procedure can be generalized to endonucleases other than Fok I.

After the oligonucleotide tags are prepared for specific hybridization, e.g. by rendering them single stranded as described above, the polynucleotides are mixed with microparticles containing the complementary sequences of the tags under conditions that favor the formation of perfectly matched duplexes between the tags and their complements. There is extensive guidance in the literature for creating these conditions. Exemplary references providing such guidance include Wetmur, Critical Reviews in Biochemistry and Molecular Biology, 26: 227-259 (1991); Sambrook et al, Molecular Cloning: A Laboratory Manual, 2nd Edition (Cold Spring Harbor Laboratory, New York, 1989); and the like. Preferably, the hybridization conditions are sufficiently stringent so that only perfectly matched sequences form stable duplexes. Under such conditions the polynucleotides specifically hybridized through their tags may be ligated to the complementary sequences attached to the microparticles. Finally, the microparticles are washed to remove polynucleotides with unligated and/or mismatched tags.

When CPG microparticles conventionally employed as synthesis supports are used, the density of tag complements on the microparticle surface is typically greater than that necessary for some sequencing operations. That is, in sequencing approaches that require successive treatment of the attached polynucleotides with a variety of enzymes, densely spaced polynucleotides may tend to inhibit access of the relatively bulky enzymes to the polynucleotides. In such cases, the polynucleotides are preferably mixed with the microparticles so that tag complements are present in significant excess, e.g. from 10:1 to 100:1, or greater, over the polynucleotides. This ensures that the density of polynucleotides on the microparticle surface will not be so high as to inhibit enzyme access. Preferably, the average inter-polynucleotide spacing on the microparticle surface is on the order of 30-100 nm. Guidance in selecting ratios for standard CPG supports and Ballotini beads (a type of solid glass support) is found in Maskos and Southern, Nucleic Acids Research, 20: 1679-1684 (1992). Preferably, for sequencing applications, standard CPG beads of diameter in the range

of 20-50 μm are loaded with about $10^5$ polynucleotides, and GMA beads of diameter in the range of 5-10 μm are loaded with a few tens of thousand of polynucleotides, e.g. $4 \times 10^4$ to $6 \times 10^4$.

In the preferred embodiment, tag complements are synthesized on microparticles combinatorially; thus, at the end of the synthesis, one obtains a complex mixture of microparticles from which a sample is taken for loading tagged polynucleotides. The size of the sample of microparticles will depend on several factors, including the size of the repertoire of tag complements, the nature of the apparatus for used for observing loaded microparticles--e.g. its capacity, the tolerance for multiple copies of microparticles with the same tag complement (i.e. "bead doubles"), and the like. The following table provide guidance regarding microparticle sample size, microparticle diameter, and the approximate physical dimensions of a packed array of microparticles of various diameters.

| Microparticle diameter | 5 μm | 10 μm | 20 μm | 40 μm |
|---|---|---|---|---|
| Max. no. polynucleotides loaded at 1 per $10^5$ sq. angstrom | | $3 \times 10^5$ | $1.26 \times 10^6$ | $5 \times 10^6$ |
| Approx. area of monolayer of $10^6$ microparticles | .45 x .45 cm | 1 x 1 cm | 2 x 2 cm | 4 x 4 cm |

The probability that the sample of microparticles contains a given tag complement or is present in multiple copies is described by the Poisson distribution, as indicated in the following table.

Table VII

| Number of microparticles in sample (as fraction of repertoire size), m | Fraction of repertoire of tag complements present in sample, $1-e^{-m}$ | Fraction of microparticles in sample with unique tag complement attached, $m(e^{-m})/2$ | Fraction of microparticles in sample carrying same tag complement as one other microparticle in sample ("bead doubles"), $m^2(e^{-m})/2$ |
|---|---|---|---|
| 1.000 | 0.63 | 0.37 | 0.18 |
| .693 | 0.50 | 0.35 | 0.12 |
| .405 | 0.33 | 0.27 | 0.05 |
| .285 | 0.25 | 0.21 | 0.03 |
| .223 | 0.20 | 0.18 | 0.02 |
| .105 | 0.10 | 0.09 | 0.005 |
| .010 | 0.01 | 0.01 | |

High Specificity Sorting and Panning

The kinetics of sorting depends on the rate of hybridization of oligonucleotide tags to their tag complements which, in turn, depends on the complexity of the tags in the hybridization reaction. Thus, a trade off exists between sorting rate and tag complexity, such that an increase in sorting rate may be achieved at the cost of reducing the complexity of the tags involved in the hybridization reaction. As explained below, the effects of this trade off may be ameliorated by "panning."

Specificity of the hybridizations may be increased by taking a sufficiently small sample so that both a high percentage of tags in the sample are unique and the nearest neighbors of substantially all the tags in a sample differ by at least two words. This latter condition may be met by taking a sample that contains a number of tag-polynucleotide conjugates that is about 0.1 percent or less of the size of the repertoire being employed. For example, if tags are constructed with eight words selected from Table II, a repertoire of $8^8$, or about $1.67 \times 10^7$, tags and tag complements are produced. In a library of tag-cDNA conjugates as described above, a 0.1 percent sample means that about 16,700 different tags are present. If this were loaded directly onto a repertoire-equivalent of microparticles, or in this example a sample of $1.67 \times 10^7$ microparticles, then only a sparse subset of the sampled microparticles would be loaded. The density of loaded microparticles can be increase--for example, for more efficient sequencing--by undertaking a "panning" step in which the sampled tag-cDNA conjugates are used to separate loaded microparticles from unloaded microparticles. Thus, in the example above, even though a "0.1 percent" sample

contains only 16,700 cDNAs, the sampling and panning steps may be repeated until as many loaded microparticles as desired are accumulated.

A panning step may be implemented by providing a sample of tag-cDNA conjugates each of which contains a capture moiety at an end opposite, or distal to, the oligonucleotide tag. Preferably, the capture moiety is of a type which can be released from the tag-cDNA conjugates, so that the tag-cDNA conjugates can be sequenced with a single-base sequencing method. Such moieties may comprise biotin, digoxigenin, or like ligands, a triplex binding region, or the like. Preferably, such a capture moiety comprises a biotin component. Biotin may be attached to tag-cDNA conjugates by a number of standard techniques. If appropriate adapters containing PCR primer binding sites are attached to tag-cDNA conjugates, biotin may be attached by using a biotinylated primer in an amplification after sampling. Alternatively, if the tag-cDNA conjugates are inserts of cloning vectors, biotin may be attached after excising the tag-cDNA conjugates by digestion with an appropriate restriction enzyme followed by isolation and filling in a protruding strand distal to the tags with a DNA polymerase in the presence of biotinylated uridine triphosphate.

After a tag-cDNA conjugate is captured, it may be released from the biotin moiety in a number of ways, such as by a chemical linkage that is cleaved by reduction, e.g. Herman et al, Anal. Biochem., 156: 48-55 (1986), or that is cleaved photochemically, e.g. Olejnik et al, Nucleic Acids Research, 24: 361-366 (1996), or that is cleaved enzymatically by introducing a restriction site in the PCR primer. The latter embodiment can be exemplified by considering the library of tag-polynucleotide conjugates described above:

```
5'-RCGACCA[C,W,W,W]9GG[T]19- cDNA -NNNR
      GGT[G,W,W,W]9CC[A]19- rDNA -NNNYCTAG-5'
```

The following adapters may be ligated to the ends of these fragments to permit amplification by PCR:

```
5'- XXXXXXXXXXXXXXXXXXX
    XXXXXXXXXXXXXXXXXXXYGAT
```

Right Adapter

```
GATCZZACTAGTZZZZZZZZZZZZ-3'
  ZZTGATCAZZZZZZZZZZZZ
```

Left Adapter

ZZ**TGATCA**ZZZZZZZZZZZZ-5'-biotin

Left Primer

where "ACTAGT" is a Spe I recognition site (which leaves a staggered cleavage ready for single base sequencing), and the X's and Z's are nucleotides selected so that the annealing and dissociation temperatures of the respective primers are approximately the same. After ligation of the adapters and amplification by PCR using the biotinylated primer, the tags of the conjugates are rendered single stranded by the exonuclease activity of T4 DNA polymerase and conjugates are combined with a sample of microparticles, e.g. a repertoire equivalent, with tag complements attached. After annealing under stringent conditions (to minimize mis-attachment of tags), the conjugates are preferably ligated to their tag complements and the loaded microparticles are separated from the unloaded microparticles by capture with avidinated magnetic beads, or like capture technique.

Returning to the example, this process results in the accumulation of about 10,500 (=16,700 x .63) loaded microparticles with different tags, which may be released from the magnetic beads by cleavage with Spe I. By repeating this process 40-50 times with new samples of microparticles and tag-cDNA conjugates, 4-5 x $10^5$ cDNAs can be accumulated by pooling the released microparticles. The pooled microparticles may then be simultaneously sequenced by a single-base sequencing technique.

Determining how many times to repeat the sampling and panning steps--or more generally, determining how many cDNAs to analyze, depends on one's objective. If the objective is to monitor the changes in abundance of relatively common sequences, e.g. making up 5% or more of a population, then relatively small samples, i.e. a small fraction of the total population size, may allow statistically significant estimates of relative abundances. On the other hand, if one seeks to monitor the abundances of rare sequences, e.g. making up 0.1% or less of a population, then large samples are required. Generally, there is a direct relationship between sample size and the reliability of the estimates of relative abundances based on the sample. There is extensive guidance in the literature on determining appropriate sample sizes for making reliable statistical estimates, e.g. Koller et al, Nucleic Acids Research, 23:185-191 (1994); Good, Biometrika, 40: 16-264 (1953); Bunge et al, J. Am. Stat. Assoc., 88: 364-373 (1993); and the like. Preferably, for

monitoring changes in gene expression based on the analysis of a series of cDNA libraries containing $10^5$ to $10^8$ independent clones of 3.0-3.5 x $10^4$ different sequences, a sample of at least $10^4$ sequences are accumulated for analysis of each library. More preferably, a sample of at least $10^5$ sequences are accumulated for the analysis of each library; and most preferably, a sample of at least 5 x $10^5$ sequences are accumulated for the analysis of each library. Alternatively, the number of sequences sampled is preferably sufficient to estimate the relative abundance of a sequence present at a frequency within the range of 0.1% to 5% with a 95% confidence limit no larger than 0.1% of the population size.

Single Base DNA Sequencing

The present invention can be employed with conventional methods of DNA sequencing, e.g. as disclosed by Hultman et al, Nucleic Acids Research, 17: 4937-4946 (1989). However, for parallel, or simultaneous, sequencing of multiple polynucleotides, a DNA sequencing methodology is preferred that requires neither electrophoretic separation of closely sized DNA fragments nor analysis of cleaved nucleotides by a separate analytical procedure, as in peptide sequencing. Preferably, the methodology permits the stepwise identification of nucleotides, usually one at a time, in a sequence through successive cycles of treatment and detection. Such methodologies are referred to herein as "single base" sequencing methods. Single base approaches are disclosed in the following references: Cheeseman, U.S. patent 5,302,509; Tsien et al, International application WO 91/06678; Rosenthal et al, International application WO 93/21340; Canard et al, Gene, 148: 1-6 (1994); and Metzker et al, Nucleic Acids Research, 22: 4259-4267 (1994).

A "single base" method of DNA sequencing which is suitable for use with the present invention and which requires no electrophoretic separation of DNA fragments is described in International application PCT/US95/03678. Briefly, the method comprises the following steps: (a) ligating a probe to an end of the polynucleotide having a protruding strand to form a ligated complex, the probe having a complementary protruding strand to that of the polynucleotide and the probe having a nuclease recognition site; (b) removing unligated probe from the ligated complex; (c) identifying one or more nucleotides in the protruding strand of the polynucleotide by the identity of the ligated probe; (d) cleaving the ligated complex with a nuclease; and (e) repeating steps (a) through (d) until the nucleotide sequence of the polynucleotide, or a portion thereof, is determined.

A single signal generating moiety, such as a single fluorescent dye, may be employed when sequencing several different target polynucleotides attached to different spatially addressable solid phase supports, such as fixed microparticles, in a

parallel sequencing operation. This may be accomplished by providing four sets of probes that are applied sequentially to the plurality of target polynucleotides on the different microparticles. An exemplary set of such probes are shown below:

| Set 1 | Set 2 | Set 3 | Set 4 |
|---|---|---|---|
| ANNNN...NN<br>N...NNTT...T* | dANNNN...NN<br>d N...NNTT...T | dANNNN...NN<br>N...NNTT...T | dANNNN...NN<br>N...NNTT...T |
| dCNNNN...NN<br>N...NNTT...T | CNNNN...NN<br>N...NNTT...T* | dCNNNN...NN<br>N...NNTT...T | dCNNNN...NN<br>N...NNTT...T |
| dGNNNN...NN<br>N...NNTT...T | dGNNNN...NN<br>N...NNTT...T | GNNNN...NN<br>N...NNTT...T* | dGNNNN...NN<br>N...NNTT...T |
| dTNNNN...NN<br>N...NNTT...T | dTNNNN...NN<br>N...NNTT...T | dTNNNN...NN<br>N...NNTT...T | TNNNN...NN<br>N...NNTT...T* |

where each of the listed probes represents a mixture of $4^3$=64 oligonucleotides such that the identity of the 3' terminal nucleotide of the top strand is fixed and the other positions in the protruding strand are filled by every 3-mer permutation of nucleotides, or complexity reducing analogs. The listed probes are also shown with a single stranded poly-T tail with a signal generating moiety attached to the terminal thymidine, shown as "T*". The "d" on the unlabeled probes designates a ligation-blocking moiety or absense of 3'-hydroxyl, which prevents unlabeled probes from being ligated. Preferably, such 3'-terminal nucleotides are dideoxynucleotides. In this embodiment, the probes of set 1 are first applied to the plurality of target polynucleotides and treated with a ligase so that target polynucleotides having a thymidine complementary to the 3' terminal adenosine of the labeled probes are ligated. The unlabeled probes are simultaneously applied to minimize inappropriate ligations. The locations of the target polynucleotides that form ligated complexes with probes terminating in "A" are identified by the signal generated by the label carried on the probe. After washing and cleavage, the probes of set 2 are applied. In this case, target polynucleotides forming ligated complexes with probes terminating in "C" are identified by location. Similarly, the probes of sets 3 and 4 are applied and locations of positive signals identified. This process of sequentially applying the four sets of probes continues until the desired number of nucleotides are identified on the target polynucleotides. Clearly, one of ordinary skill could construct similar sets of probes that could have many variations, such as having protruding strands of different lengths, different moieties to block ligation of unlabeled probes, different means for labeling probes, and the like.

Apparatus for Sequencing Populations of Polynucleotides

An objective of the invention is to sort identical molecules, particularly polynucleotides, onto the surfaces of microparticles by the specific hybridization of tags and their complements. Once such sorting has taken place, the presence of the molecules or operations performed on them can be detected in a number of ways depending on the nature of the tagged molecule, whether microparticles are detected separately or in "batches," whether repeated measurements are desired, and the like. Typically, the sorted molecules are exposed to ligands for binding, e.g. in drug development, or are subjected chemical or enzymatic processes, e.g. in polynucleotide sequencing. In both of these uses it is often desirable to simultaneously observe signals corresponding to such events or processes on large numbers of microparticles. Microparticles carrying sorted molecules (referred to herein as "loaded" microparticles) lend themselves to such large scale parallel operations, e.g. as demonstrated by Lam et al (cited above).

Preferably, whenever light-generating signals, e.g. chemiluminescent, fluorescent, or the like, are employed to detect events or processes, loaded microparticles are spread on a planar substrate, e.g. a glass slide, for examination with a scanning system, such as described in International patent applications PCT/US91/09217, PCT/NL90/00081, and PCT/US95/01886. The scanning system should be able to reproducibly scan the substrate and to define the positions of each microparticle in a predetermined region by way of a coordinate system. In polynucleotide sequencing applications, it is important that the positional identification of microparticles be repeatable in successive scan steps.

Such scanning systems may be constructed from commercially available components, e.g. x-y translation table controlled by a digital computer used with a detection system comprising one or more photomultiplier tubes, or alternatively, a CCD array, and appropriate optics, e.g. for exciting, collecting, and sorting fluorescent signals. In some embodiments a confocal optical system may be desirable. An exemplary scanning system suitable for use in four-color sequencing is illustrated diagrammatically in Figure 5. Substrate 300, e.g. a microscope slide with fixed microparticles, is placed on x-y translation table 302, which is connected to and controlled by an appropriately programmed digital computer 304 which may be any of a variety of commercially available personal computers, e.g. 486-based machines or PowerPC model 7100 or 8100 available form Apple Computer (Cupertino, CA). Computer software for table translation and data collection functions can be provided by commercially available laboratory software, such as Lab Windows, available from National Instruments.

Substrate 300 and table 302 are operationally associated with microscope 306 having one or more objective lenses 308 which are capable of collecting and delivering light to microparticles fixed to substrate 300. Excitation beam 310 from light source 312, which is preferably a laser, is directed to beam splitter 314, e.g. a dichroic mirror, which re-directs the beam through microscope 306 and objective lens 308 which, in turn, focuses the beam onto substrate 300. Lens 308 collects fluorescence 316 emitted from the microparticles and directs it through beam splitter 314 to signal distribution optics 318 which, in turn, directs fluorescence to one or more suitable opto-electronic devices for converting some fluorescence characteristic, e.g. intensity, lifetime, or the like, to an electrical signal. Signal distribution optics 318 may comprise a variety of components standard in the art, such as bandpass filters, fiber optics, rotating mirrors, fixed position mirrors and lenses, diffraction gratings, and the like. As illustrated in Figure 2, signal distribution optics 318 directs fluorescence 316 to four separate photomultiplier tubes, 330, 332, 334, and 336, whose output is then directed to pre-amps and photon counters 350, 352, 354, and 356. The output of the photon counters is collected by computer 304, where it can be stored, analyzed, and viewed on video 360. Alternatively, signal distribution optics 318 could be a diffraction grating which directs fluorescent signal 318 onto a CCD array.

The stability and reproducibility of the positional localization in scanning will determine, to a large extent, the resolution for separating closely spaced microparticles. Preferably, the scanning systems should be capable of resolving closely spaced microparticles, e.g. separated by a particle diameter or less. Thus, for most applications, e.g. using CPG microparticles, the scanning system should at least have the capability of resolving objects on the order of 10-100 μm. Even higher resolution may be desirable in some embodiments, but with increase resolution, the time required to fully scan a substrate will increase: thus, in some embodiments a compromise may have to be made between speed and resolution. Increases in scanning time can be achieved by a system which only scans positions where microparticles are known to be located, e.g from an initial full scan. Preferably, microparticle size and scanning system resolution are selected to permit resolution of fluorescently labeled microparticles randomly disposed on a plane at a density between about ten thousand to one hundred thousand microparticles per $cm^2$.

In sequencing applications, loaded microparticles can be fixed to the surface of a substrate in variety of ways. The fixation should be strong enough to allow the microparticles to undergo successive cycles of reagent exposure and washing without significant loss. When the substrate is glass, its surface may be derivatized with an alkylamino linker using commercially available reagents, e.g. Pierce Chemical, which

in turn may be cross-linked to avidin, again using conventional chemistries, to form an avidinated surface. Biotin moieties can be introduced to the loaded microparticles in a number of ways. For example, a fraction, e.g. 10-15 percent, of the cloning vectors used to attach tags to polynucleotides are engineered to contain a unique restriction site (providing sticky ends on digestion) immediately adjacent to the polynucleotide insert at an end of the polynucleotide opposite of the tag. The site is excised with the polynucleotide and tag for loading onto microparticles. After loading, about 10-15 percent of the loaded polynucleotides will possess the unique restriction site distal from the microparticle surface. After digestion with the associated restriction endonuclease, an appropriate double stranded adaptor containing a biotin moiety is ligated to the sticky end. The resulting microparticles are then spread on the avidinated glass surface where they become fixed via the biotin-avidin linkages.

Alternatively and preferably when sequencing by ligation is employed, in the initial ligation step a mixture of probes is applied to the loaded microparticle: a fraction of the probes contain a type IIs restriction recognition site, as required by the sequencing method, and a fraction of the probes have no such recognition site, but instead contain a biotin moiety at its non-ligating end. Preferably, the mixture comprises about 10-15 percent of the biotinylated probe.

In still another alternative, when DNA-loaded microparticles are applied to a glass substrate, the DNA may nonspecifically adsorb to the glass surface upon several hours, e.g. 24 hours, incubation to create a bond sufficiently strong to permit repeated exposures to reagents and washes without significant loss of microparticles. Preferably, such a glass substrate is a flow cell, which may comprise a channel etched in a glass slide. Preferably, such a channel is closed so that fluids may be pumped through it and has a depth sufficiently close to the diameter of the microparticles so that a monolayer of microparticles is trapped within a defined observation region.

## Identification of Novel Polynucleotides in cDNA Libraries

Novel polynucleotides in a cDNA library can be identified by constructing a library of cDNA molecules attached to microparticles, as described above. A large fraction of the library, or even the entire library, can then be partially sequenced in parallel. After isolation of mRNA, and perhaps normalization of the population as taught by Soares et al, Proc. Natl. Acad. Sci., 91: 9228-9232 (1994), or like references, the following primer may by hybridized to the polyA tails for first strand synthesis with a reverse transcriptase using conventional protocols (SEQ ID NO: 1):

```
5'-mRNA- [A]n -3'
         [T]19-[primer site]-GG[W,W,W,C]9ACCAGCTGATC-5'
```

where $[W,W,W,C]_9$ represents a tag as described above, "ACCAGCTGATC" is an optional sequence forming a restriction site in double stranded form, and "primer site" is a sequence common to all members of the library that is later used as a primer binding site for amplifying polynucleotides of interest by PCR.

After reverse transcription and second strand synthesis by conventional techniques, the double stranded fragments are inserted into a cloning vector as described above and amplified. The amplified library is then sampled and the sample amplified. The cloning vectors from the amplified sample are isolated, and the tagged cDNA fragments excised and purified. After rendering the tag single stranded with a polymerase as described above, the fragments are methylated and sorted onto microparticles in accordance with the invention. Preferably, as described above, the cloning vector is constructed so that the tagged cDNAs can be excised with an endonuclease, such as Fok I, that will allow immediate sequencing by the preferred single base method after sorting and ligation to microparticles.

Stepwise sequencing is then carried out simultaneously on the whole library, or one or more large fractions of the library, in accordance with the invention until a sufficient number of nucleotides are identified on each cDNA for unique representation in the genome of the organism from which the library is derived. For example, if the library is derived from mammalian mRNA then a randomly selected sequence 14-15 nucleotides long is expected to have unique representation among the 2-3 thousand megabases of the typical mammalian genome. Of course identification of far fewer nucleotides would be sufficient for unique representation in a library derived from bacteria, or other lower organisms. Preferably, at least 20-30 nucleotides are identified to ensure unique representation and to permit construction of a suitable primer as described below. The tabulated sequences may then be compared to known sequences to identify unique cDNAs.

Unique cDNAs are then isolated by conventional techniques, e.g. constructing a probe from the PCR amplicon produced with primers directed to the prime site and the portion of the cDNA whose sequence was determined. The probe may then be used to identify the cDNA in a library using a conventional screening protocol.

The above method for identifying new cDNAs may also be used to fingerprint mRNA populations, either in isolated measurements or in the context of a dynamically changing population. Partial sequence information is obtained simultaneously from a large sample, e.g. ten to a hundred thousand, or more, of cDNAs attached to separate microparticles as described in the above method.

**Example 1**

Construction of a Tag Library

An exemplary tag library is constructed as follows to form the chemically synthesized 9-word tags of nucleotides A, G, and T defined by the formula:

$$3'\text{-TGGC-}[^4(A,G,T)_9]\text{-CCCCp}$$

where "$[^4(A,G,T)_9]$" indicates a tag mixture where each tag consists of nine 4-mer words of A, G, and T; and "p" indicate a 5' phosphate. This mixture is ligated to the following right and left primer binding regions (SEQ ID NO: 4 and SEQ ID NO 5):

```
5'- AGTGGCTGGGCATCGGACCG          5'- GGGGCCCAGTCAGCGTCGAT
    TCACCGACCCGTAGCCp                     GGGTCAGTCGCAGCTA
```

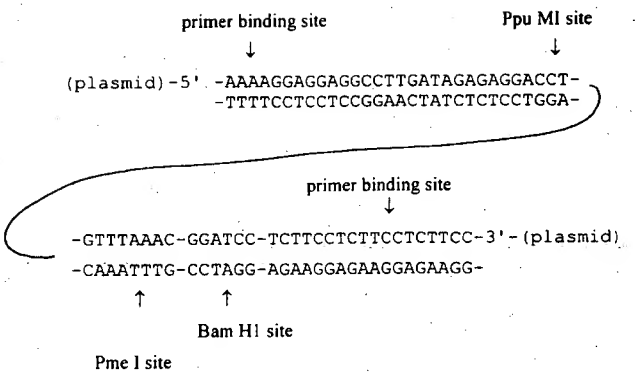
LEFT RIGHT

The right and left primer binding regions are ligated to the above tag mixture, after which the single stranded portion of the ligated structure is filled with DNA polymerase then mixed with the right and left primers indicated below and amplified to give a tag library (SEQ ID NO: 6).

```
      Left Primer

5'- AGTGGCTGGGCATCGGACCG

5'- AGTGGCTGGGCATCGGACCG-  [4(A,G,T)9]-GGGGCCCAGTCAGCGTCGAT
    TCACCGACCCGTAGCCTGGC-  [4(A,G,T)9]-CCCCGGGTCAGTCGCAGCTA

                                        CCCCGGGTCAGTCGCAGCTA-5'

                                        Right Primer
```

The underlined portion of the left primer binding region indicates a Rsr II recognition site. The left-most underlined region of the right primer binding region indicates recognition sites for Bsp 120I, Apa I, and Eco O 109I, and a cleavage site for Hga I. The right-most underlined region of the right primer binding region indicates the recognition site for Hga I. Optionally, the right or left primers may be synthesized with a biotin attached (using conventional reagents, e.g. available from Clontech Laboratories, Palo Alto, CA) to facilitate purification after amplification and/or cleavage.

NOT FURNISHED UPON FILING




The plasmid is cleaved with Ppu MI and Pme I (to give a Rsr II-compatible end and a flush end so that the insert is oriented) and then methylated with DAM methylase. The tag-containing construct is cleaved with Rsr II and then ligated to the open plasmid, after which the conjugate is cleaved with Mbo I and Bam HI to permit ligation and closing of the plasmid. The plasmid is then amplified and isolated and used in accordance with the invention.

**Example 3**

Changes in Gene Expression Profiles in Liver Tissue of Rats Exposed to Various Xenobiotic Agents

In this experiment, to test the capability of the method of the invention to detect genes induced as a result of exposure to xenobiotic compounds, the gene expression profile of rat liver tissue is examined following administration of several compounds known to induce the expression of cytochrome P-450 isoenzymes. The results obtained from the method of the invention are compared to results obtained from reverse transcriptase PCR measurements and immunochemical measurements of the cytochrome P-450 isoenzymes. Protocols and materials for the latter assays are described in Morris et al, Biochemical Pharmacology, 52: 781-792 (1996).

Male Sprague-Dawley rats between the ages of 6 and 8 weeks and weighing 200-300 g are used, and food and water are available to the animals *ad lib*. Test compounds are phenobarbital (PB), metyrapone (MET), dexamethasone (DEX), clofibrate (CLO), corn oil (CO), and β-naphthoflavone (BNF), and are available from Sigma Chemical Co. (St. Louis, MO). Antibodies against specific P-450 enzymes are available from the following sources: rabbit anti-rat CYP3A1 from Human Biologics, Inc. (Phoenix, AZ); goat anti-rat CYP4A1 from Daiichi Pure Chemicals Co. (Tokyo,

Japan); monoclonal mouse anti-rat CYP1A1, monoclonal mouse anti-rat CYP2C11, goat anti-rat CYP2E1, and monoclonal mouse anti-rat CYP2B1 from Oxford Biochemical Research, Inc. (Oxford, MI). Secondary antibodies (goat anti-rabbit IgG, rabbit anti-goat IgG and goat anti-mouse IgG) are available from Jackson ImmunoResearch Laboratories (West Grove, PA).

Animals are administered either PB (100 mg/kg), BNF (100 mg/kg), MET (100 mg/kg), DEX (100 mg/kg), or CLO (250 mg/kg) for 4 consecutive days via intraperitoneal injection following a dosing regimen similar to that described by Wang et al, Arch. Biochem. Biophys. 290: 355-361 (1991). Animals treated with $H_2O$ and CO are used as controls. Two hours following the last injection (day 4), animals are killed, and the livers are removed. Livers are immediately frozen and stored at -70°C.

Total RNA is prepared from frozen liver tissue using a modification of the method described by Xie et al, Biotechniques, 11: 326-327 (1991). Approximately 100-200 mg of liver tissue is homogenized in the RNA extraction buffer described by Xie et al to isolate total RNA. The resulting RNA is reconstituted in diethylpyrocarbonate-treated water, quantified spectrophotometrically at 260 nm, and adjusted to a concentration of 100 µg/ml. Total RNA is stored in diethylpyrocarbonate-treated water for up to 1 year at -70°C without any apparent degradation. RT-PCR and sequencing are performed on samples from these preparations.

For sequencing, samples of RNA corresponding to about 0.5 µg of poly(A)$^+$ RNA are used to construct libraries of tag-cDNA conjugates following the protocol described in the section entitled "Attaching Tags to Polynucleotides for Sorting onto Solid Phase Supports," with the following exception: the tag repertoire is constructed from six 4-nucleotide words from Table II. Thus, the complexity of the repertoire is $8^6$ or about $2.6 \times 10^5$. For each tag-cDNA conjugate library constructed, ten samples of about ten thousand clones are taken for amplification and sorting. Each of the amplified samples is separately applied to a fixed monolayer of about $10^6$ 10 µm diameter GMA beads containing tag complements. That is, the "sample" of tag complements in the GMA bead population on each monolayer is about four fold the total size of the repertoire, thus ensuring there is a high probability that each of the sampled tag-cDNA conjugates will find its tag complement on the monolayer. After the oligonucleotide tags of the amplified samples are rendered single stranded as described above, the tag-cDNA conjugates of the samples are separately applied to the monolayers under conditions that permit specific hybridization only between oligonucleotide tags and tag complements forming perfectly matched duplexes. Concentrations of the amplified samples and hybridization times are selected to

permit the loading of about 5 x $10^4$ to 2 x $10^5$ tag-cDNA conjugates on each bead where perfect matches occur. After ligation, 9-12 nucleotide portions of the attached cDNAs are determined in parallel by the single base sequencing technique described by Brenner in International patent application PCT/US95/03678. Frequency distributions for the gene expression profiles are assembled from the sequence information obtained from each of the ten samples.

RT-PCRs of selected mRNAs corresponding to cytochrome P-450 genes and the constitutively expressed cyclophilin gene are carried out as described in Morris et al (cited above). Briefly, a 20 μL reaction mixture is prepared containing 1x reverse transcriptase buffer (Gibco BRL), 10 nM dithiothreitol, 0.5 nM dNTPs, 2.5 μM oligo $d(T)_{15}$ primer, 40 units RNasin (Promega, Madison, WI), 200 units RNase H-reverse transcriptase (Gibco BRL), and 400 ng of total RNA (in diethylpyrocarbonate-treated water). The reaction is incubated for 1 hour at 37°C followed by inactivation of the enzyme at 95°C for 5 min. The resulting cDNA is stored at -20°C until used. For PCR amplification of cDNA, a 10 μL reaction mixture is prepared containing 10x polymerase reaction buffer, 2 mM $MgCl_2$, 1 unit Taq DNA polymerase (Perkin-Elmer, Norwalk, CT), 20 ng cDNA, and 200 nM concentration of the 5' and 3' specific PCR primers of the sequences described in Morris et al (cited above). PCRs are carried out in a Perkin-Elmer 9600 thermal cycler for 23 cycles using melting, annealing, and extension conditions of 94°C for 30 sec., 56°C for 1 min., and 72°C for 1 min., respectively. Amplified cDNA products are separated by PAGE using 5% native gels. Bands are detected by staining with ethidium bromide.

Western blots of the liver proteins are carried out using standard protocols after separation by SDS-PAGE. Briefly, proteins are separated on 10% SDS-PAGE gels under reducing conditions and immunoblotted for detection of P-450 isoenzymes using a modification of the methods described in Harris et al, Proc. Natl. Acad. Sci., 88: 1407-1410 (1991). Protein are loaded at 50 μg/lane and resolved under constant current (250 V) for approximately 4 hours at 2°C. Proteins are transferred to nitrocellulose membranes (Bio-Rad, Hercules, CA) in 15 mM Tris buffer containing 120 mM glycine and 20% (v/v) methanol. The nitrocellulose membranes are blocked with 2.5% BSA and immunoblotted for P-450 isoenzymes using primary monoclonal and polyclonal antibodies and secondary alkaline phosphatase conjugated anti-IgG. Immunoblots are developed with the Bio-Rad alkaline phosphatase substrate kit.

The three types of measurements of P-450 isoenzyme induction showed substantial agreement.

## APPENDIX Ia

Exemplary computer program for generating minimally cross hybridizing sets

(single stranded tag/single stranded tag complement)

```
Program minxh
c
c
c
      integer*2 sub1(6),mset1(1000,6),mset2(1000,6)
      dimension nbase(6)
c
c
      write(*,*)'ENTER SUBUNIT LENGTH'
      read(*,100)nsub
100   format(i1)
      open(1,file='sub4.dat',form='formatted',status='new')
c
c
        nset=0
        do 7000 m1=1,3
           do 7000 m2=1,3
              do 7000 m3=1,3
                 do 7000 m4=1,3
                    sub1(1)=m1
                    sub1(2)=m2
                    sub1(3)=m3
                    sub1(4)=m4
c
c
      ndiff=3
c
c
c                 Generate set of subunits differing from
c                 sub1 by at least ndiff nucleotides.
c                 Save in mset1.
c
c
        jj=1
        do 900 j=1,nsub
900        mset1(1,j)=sub1(j)
c
c
        do 1000 k1=1,3
           do 1000 k2=1,3
              do 1000 k3=1,3
                 do 1000 k4=1,3
c
c
                 nbase(1)=k1
                 nbase(2)=k2
                 nbase(3)=k3
                 nbase(4)=k4
```

```
c
                n=0
                do 1200 j=1,nsub
                   if(subl(j).eq.1 .and. nbase(j).ne.1 .or.
     1                subl(j).eq.2 .and. nbase(j).ne.2 .or.
     3                subl(j).eq.3 .and. nbase(j).ne.3) then
                      n=n+1
                      endif
1200                  continue
c
c
          if(n.ge.ndiff) then
c
c
c                                        If number of mismatches
c                                        is greater than or equal
c                                        to ndiff then record
c                                        subunit in matrix mset
c
c
               jj=jj+1
                do 1100 i=1,nsub
1100               mset1(jj,i)=nbase(i)
               endif
c
c
1000       continue
c
c
               do 1325 j2=1,nsub
               mset2(1,j2)=mset1(1,j2)
1325           mset2(2,j2)=mset1(2,j2)
c
c
c                                        Compare subunit 2 from
c                                        mset1 with each successive
c                                        subunit in mset1, i.e. 3,
c                                        4,5, ... etc. Save those
c                                        with mismatches .ge. ndiff
c                                        in matrix mset2 starting at
c                                        position 2.
c                                          Next transfer contents
c                                        of mset2 into mset1 and
c                                        start
c                                        comparisons again this time
c                                        starting with subunit 3.
c                                        Continue until all subunits
c                                        undergo the comparisons.
c
c
          npass=0
c
c
1700      continue
          kk=npass+2
          npass=npass+1
c
```

```
c
          do 1500 m=npass+2,jj
             n=0
             do 1600 j=1,nsub
                if(mset1(npass+1,j).eq.1.and.mset1(m,j).ne.1.or.
     2             mset1(npass+1,j).eq.2.and.mset1(m,j).ne.2.or.
     2             mset1(npass+1,j).eq.3.and.mset1(m,j).ne.3) then
                   n=n+1
                   endif
1600               continue
                if(n.ge.ndiff) then
                      kk=kk+1
                      do 1625 i=1,nsub
1625                     mset2(kk,i)=mset1(m,i)
                endif
1500            continue
c
c                                        kk is the number of subunits
c                                        stored in mset2
c
c                                        Transfer contents of mset2
c                                        into mset1 for next pass.
c
c
              do 2000 k=1,kk
                 do 2000 m=1,nsub
2000                mset1(k,m)=mset2(k,m)
          if(kk.lt.jj) then
             jj=kk
             goto 1700
             endif
c
c
           nset=nset+1
          write(1,7009)
7009       format(/)
          do 7008 k=1,kk
7008          write(1,7010)(mset1(k,m),m=1,nsub)
7010      format(4i1)
          write(*,*)
          write(*,120) kk,nset
120       format(1x,'Subunits in set=',i5,2x,'Set No=',i5)
7000        continue
           close(1)
c
c
          end
c                 ********************************
c                 ********************************
```

# APPENDIX Ib

Exemplary computer program for generating minimally cross hybridizing sets

(single stranded tag/single stranded tag complement)

```
      Program tagN
c
c
c           Program tagN generates minimally cross-hybridizing
c             sets of subunits given i) N--subunit length, and ii)
c             an initial subunit sequence.  tagN assumes that only
c             3 of the four natural nucleotides are used in the tags.
c
c
c
            character*1 subl(20)
            integer*2 mset(10000,20), nbase(20)
c
c
            write(*,*)'ENTER SUBUNIT LENGTH'
            read(*,100)nsub
  100       format(i2)
c
c
            write(*,*)'ENTER SUBUNIT SEQUENCE'
            read(*,110)(subl(k),k=1,nsub)
  110       format(20a1)
c
c
            ndiff=10
c
c
c                     Let a=1 c=2 g=3 & t=4
c
c
c
              do 800 kk=1,nsub
              if(subl(kk).eq.'a') then
                 mset(1,kk)=1
                 endif
                      if(subl(kk).eq.'c') then
                         mset(1,kk)=2
                         endif
                              if(subl(kk).eq.'g') then
                                 mset(1,kk)=3
                                 endif
                                      if(subl(kk).eq.'t') then
                                         mset(1,kk)=4
                                         endif
  800     continue
c
c
c                     Generate set of subunits differing from
c                     subl by at least ndiff nucleotides.
c
c
c
              jj=1
c
c
              do 1000 k1=1,3
```

```
                do 1000 k2=1,3
                  do 1000 k3=1,3
                    do 1000 k4=1,3
                      do 1000 k5=1,3
                        do 1000 k6=1,3
                          do 1000 k7=1,3
                            do 1000 k8=1,3
                              do 1000 k9=1,3
                                do 1000 k10=1,3
              do 1000 k11=1,3
                do 1000 k12=1,3
                  do 1000 k13=1,3
                    do 1000 k14=1,3
                      do 1000 k15=1,3
                        do 1000 k16=1,3
                          do 1000 k17=1,3
                            do 1000 k18=1,3
                              do 1000 k19=1,3
                                do 1000 k20=1,3
c
c
                     nbase(1)=k1
                     nbase(2)=k2
                     nbase(3)=k3
                     nbase(4)=k4
                     nbase(5)=k5
                     nbase(6)=k6
                     nbase(7)=k7
                     nbase(8)=k8
                     nbase(9)=k9
                     nbase(10)=k10
                     nbase(11)=k11
                     nbase(12)=k12
                     nbase(13)=k13
                     nbase(14)=k14
                     nbase(15)=k15
                     nbase(16)=k16
                     nbase(17)=k17
                     nbase(18)=k18
                     nbase(19)=k19
                     nbase(20)=k20
c
c
                  do 1250 nn=1,jj
c
                   n=0
                   do 1200 j=1,nsub
                      if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
     1                   mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
     2                   mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
     3                   mset(nn,j).eq.4 .and. nbase(j).ne.4) then
                      n=n+1
                      endif
1200                  continue
c
c
          if(n.lt.ndiff) then
             goto 1000
             endif
1250      continue
c
c
               jj=jj+1
               write(*,130)(nbase(i),i=1,nsub),jj
                do 1100 i=1,nsub
```

```
                    mset(jj,i)=nbase(i)
1100                   continue
c
c
1000      continue
c
c
          write(*,*)
130       format(10x,20(1x,i1),5x,i5)
          write(*,*)
          write(*,120) jj
120       format(1x,'Number of words=',i5)
c
c
          end
c
c               ******************************************
c               ******************************************
```

# APPENDIX Ic

## Exemplary computer program for generating minimally cross hybridizing sets

## (double stranded tag/single stranded tag complement)

```
Program 3tagN
c
c
c          Program 3tagN generates minimally cross-hybridizing
c            sets of duplex subunits given i) N--subunit length,
c            and ii) an initial homopurine sequence.
c
c
        character*1 subl(20)
        integer*2 mset(10000,20), nbase(20)
c
c
        write(*,*)'ENTER SUBUNIT LENGTH'
        read(*,100)nsub
100     format(i2)
c
c
        write(*,*)'ENTER SUBUNIT SEQUENCE a & g only'
        read(*,110)(subl(k),k=1,nsub)
110     format(20a1)
c
        ndiff=10
c
c
c                   Let a=1 and g=2

         do 800 kk=1,nsub
         if(subl(kk).eq.'a') then
            mset(1,kk)=1
            endif
                 if(subl(kk).eq.'g') then
                    mset(1,kk)=2
                    endif
800     continue
c


         jj=1
c
         do 1000 k1=1,3
           do 1000 k2=1,3
             do 1000 k3=1,3
               do 1000 k4=1,3
                 do 1000 k5=1,3
                   do 1000 k6=1,3
                     do 1000 k7=1,3
                       do 1000 k8=1,3
                         do 1000 k9=1,3
                           do 1000 k10=1,3
         do 1000 k11=1,3
           do 1000 k12=1,3
             do 1000 k13=1,3
               do 1000 k14=1,3
                 do 1000 k15=1,3
                   do 1000 k16=1,3
                     do 1000 k17=1,3
                       do 1000 k18=1,3
```

```
                                  do 1000 k19=1,3
                                    do 1000 k20=1,3
c
                        nbase(1)=k1
                        nbase(2)=k2
                        nbase(3)=k3
                        nbase(4)=k4
                        nbase(5)=k5
                        nbase(6)=k6
                        nbase(7)=k7
                        nbase(8)=k8
                        nbase(9)=k9
                        nbase(10)=k10
                        nbase(11)=k11
                        nbase(12)=k12
                        nbase(13)=k13
                        nbase(14)=k14
                        nbase(15)=k15
                        nbase(16)=k16
                        nbase(17)=k17
                        nbase(18)=k18
                        nbase(19)=k19
                        nbase(20)=k20
c
                    do 1250 nn=1,jj
c
                     n=0
                     do 1200 j=1,nsub
                        if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
     1                    mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
     2                    mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
     3                    mset(nn,j).eq.4 .and. nbase(j).ne.4) then
                          n=n+1
                          endif
1200                      continue
c
              if(n.lt.ndiff) then
                 goto 1000
                 endif
1250          continue
c
                   jj=jj+1
                   write(*,130)(nbase(i),i=1,nsub),jj
                    do 1100 i=1,nsub
                       mset(jj,i)=nbase(i)
1100                      continue
c
1000          continue
c
              write(*,*)
130           format(10x,20(1x,i1),5x,i5)
              write(*,*)
              write(*,120) jj
120           format(1x,'Number of words=',i5)
c
c
              end
```

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: David W. Martin, Jr.

(ii) TITLE OF INVENTION: Measurement of Gene Expression profiles in Toxicity Determination

(iii) NUMBER OF SEQUENCES: 7

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: Stephen C. Macevicz, Lynx Therapeutics, Inc.
(B) STREET: 3832 Bay Center Place
(C) CITY: Hayward
(D) STATE: California
(E) COUNTRY: USA
(F) ZIP: 94545

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: 3.5 inch diskette
(B) COMPUTER: IBM compatible
(C) OPERATING SYSTEM: Windows 3.1
(D) SOFTWARE: Microsoft Word 5.1

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER:
(B) FILING DATE:
(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:
(A) APPLICATION NUMBER: PCT/US96/09513
(B) FILING DATE: 06-JUN-96

(vii) PRIOR APPLICATION DATA:
(A) APPLICATION NUMBER: PCT/US95/12791
(B) FILING DATE: 12-OCT-95

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: Stephen C. Macevicz
(B) REGISTRATION NUMBER: 30,285
(C) REFERENCE/DOCKET NUMBER: 813wo

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: (510) 670-9365
(B) TELEFAX: (510) 670-9302

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

CTAGTCGACC A 11

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 11 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

NRRGATCYNN N 11

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 38 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GAGGATGCCT TTATGGATCC ACTCGAGATC CCAATCCA 38

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

AGTGGCTGGG CATCGGACCG 20

(2) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 nucleotides
(B) TYPE: nucleic acid

(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

GGGGCCCAGT CAGCGTCGAT 20

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

ATCGACGCTG ACTGGGCCCC 16

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 62 nucleotides
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

AAAAGGAGGA GGCCTTGATA GAGAGGACCT GTTTAAACGG ATCCTCTTCC 50
TCTTCCTCTT CC 62

I claim:

1. A method of determining the toxicity of a compound, the method comprising the steps of:

administering the compound to a test organism;

extracting a population of mRNA molecules from each of one or more tissues of the test organism;

forming a separate population of cDNA molecules from each population of mRNA molecules from the one or more tissues such that each cDNA molecule of a separate population has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set;

separately sampling each population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached;

sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports;

determining the nucleotide sequence of a portion of each of the sorted cDNA molecules of each separate population to form a frequency distribution of expressed genes for each of the one or more tissues; and

correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

2. The method of claim 1 wherein said oligonucleotide tag and said complement of said oligonucleotide tag are single stranded.

3. The method of claim 2 wherein said oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length and each subunit being selected from the same minimally cross-hybridizing set.

4. The method of claim 3 wherein said one or more solid phase supports are microparticles and wherein said step of sorting said cDNA molecules onto the microparticles produces a subpopulation of loaded microparticles and a subpopulation of unloaded microparticles.

5. The method of claim 4 further including a step of separating said loaded microparticles from said unloaded microparticles.

6. The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.

7. The method of claim 6 wherein said number of loaded microparticles is at least 100,000.

8. The method of claim 7 wherein said number of loaded microparticles is at least 500,000.

9. The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.

10. The method of claim 4 wherein said test organism is a mammalian tissue culture.

11. The method of claim 10 wherein said mammalian tissue culture comprises hepatocytes.

12. The method of claim 4 wherein said test organism is an animal selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.

13. The method of claim 12 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancrease urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.

14. A method of identifying genes which are differentially expressed in a selected tissue of a test animal after treatment with a compound, the method comprising the steps of:

administering the compound to a test animal;

extracting a population of mRNA molecules from the selected tissue of the test animal;

forming a population of cDNA molecules from the population of mRNA molecules such that each cDNA molecule has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set;

sampling the population of cDNA molecules such that substantially all different cDNA molecules have different oligonucleotide tags attached;

sorting the cDNA molecules by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports;

determining the nucleotide sequence of a portion of each of the sorted cDNA molecules to form a frequency distribution of expressed genes; and

identifying genes expressed in response to administering the compound by comparing the frequencing distribution of expressed genes of the selected tissue of the test animal with a frequency distribution of expressed genes of the selected tissue of a control animal.

15. The method of claim 14 wherein said oligonucleotide tag and said complement of said oligonucleotide tag are single stranded.

16. The method of claim 15 wherein said oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length and each subunit being selected from the same minimally cross-hybridizing set.

17. The method of claim 16 wherein said one or more solid phase supports are microparticles and wherein said step of sorting said cDNA molecules onto the microparticles produces a subpopulation of loaded microparticles and a subpopulation of unloaded microparticles.

18. The method of claim 17 further including a step of separating said loaded microparticles from said unloaded microparticles.

19. The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.

20. The method of claim 19 wherein said number of loaded microparticles is at least 100,000.

21. The method of claim 20 wherein said number of loaded microparticles is at least 500,000.

22. The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.

23. The method of claim 17 wherein said test animal is selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.

24. The method of claim 23 wherein said selected tissue is selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancrease urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.

25. A use of the technique of massively parallel signature sequencing to determine the toxicity of a compound in a test organism, the use comprising the steps of:

administering the compound to a test organism;

extracting a population of mRNA molecules from each of one or more tissues of the test organism and forming a population of cDNA molecules for each of the one or more tissues;

determining the nucleotide sequence of a portion of each of the cDNA molecules of each separate population using massively parallel signature sequencing to form a frequency distribution of expressed genes for each of the one or more tissues; and

correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

26. The use of claim 25 wherein said test organism is a mammalian tissue culture.

27. The use of claim 26 wherein said mammalian tissue culture comprises hepatocytes.

28. The use of claim 25 wherein said test organism is an animal selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.

29. The use of claim 28 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancrease urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.

30. A use of the technique of massively parallel signature sequencing to identify genes which are differentially expressed in a test organism after treatment with a compound and which are correlated with toxicity of the compound, the use comprising the steps of:

administering the compound to the test organism;

extracting a population of mRNA molecules from a selected tissue of the test organism and forming a population of cDNA molecules;

determining the nucleotide sequence of a portion of each of the cDNA molecules using massively parallel signature sequencing to form a frequency distribution of expressed genes;

identifying genes expressed in response to administering the compound by comparing the frequencing distribution of expressed genes of the selected tissue of the test organism with a frequency distribution of expressed genes of the selected tissue of a control organism; and

determining whether the genes expressed in response to administering the compound are correlated with toxicity of the compound in the test organism.

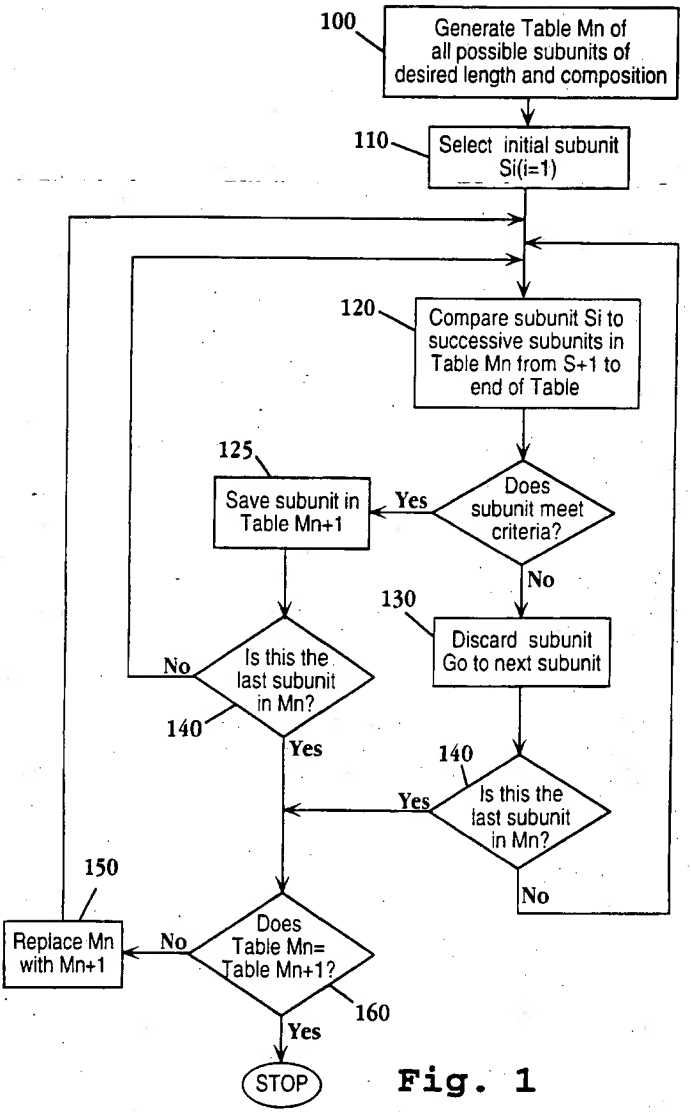

100
Generate Table Mn of all possible subunits of desired length and composition
110
Select initial subunit Si(i=1)
120
Compare subunit Si to successive subunits in Table Mn from S+1 to end of Table
125
Save subunit in Table Mn+1
Yes
Does subunit meet criteria?
No
130
Discard subunit Go to next subunit
No
Is this the last subunit in Mn?
140
Yes
140
Yes
Is this the last subunit in Mn?
No
150
Replace Mn with Mn+1
No
Does Table Mn= Table Mn+1?
160
Yes
STOP

Fig. 1

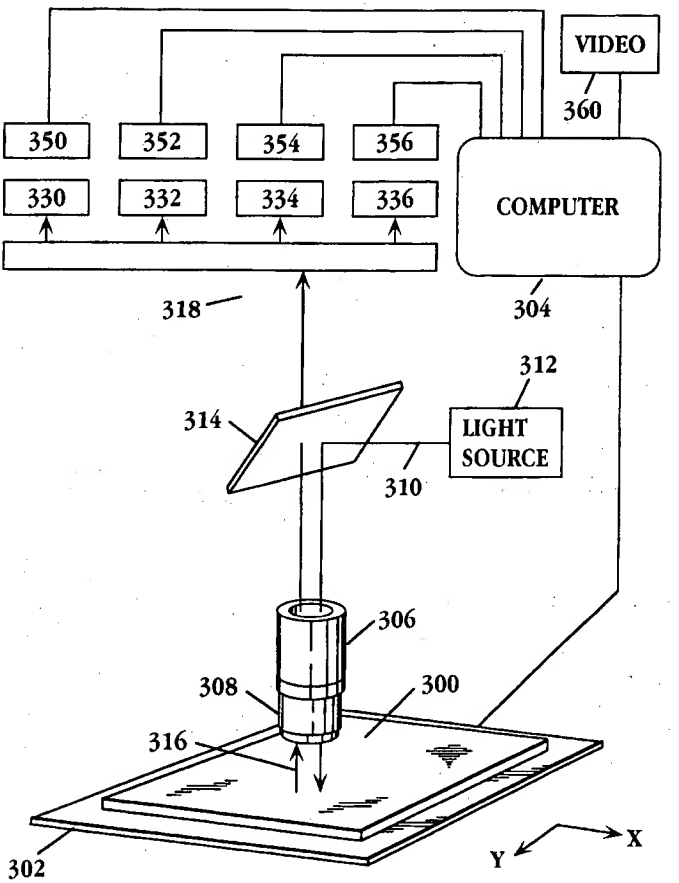

VIDEO
360
350
352
354
356
330
332
334
336
COMPUTER
318
304
312
314
LIGHT SOURCE
310
306
308
300
316
302
Y
X

Fig. 2

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/16342

**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(6) : C12Q 1/68; C07H 21/04
US CL : 435/6; 536/24.3
According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 536/24.3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, MEDLINE, BIOSIS, CAPLUS, SCISEARCH
search terms: Martin, David W., toxic?, differential?, express?, cDNA, mRNA, RNA, gene#, hybrid?,

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A | CHETVERIN et al. Oligonucleotide arrays: New concepts and possibilities. Bio/Technology. 12 November 1994, Vol. 12, pages 1093-1099, especially pages 1095-1096. | 1-30 |
| A | BRENNER et al. Encoded combinatorial chemistry. Proceedings of the National Academy of Sciences USA. June 1992, Vol. 89, pages 5381-5383. | 1-30 |
| A | MATSUBARA et al. cDNA analyses in the human genome project. Gene. 15 December 1993, Vol. 135, No. 1-2, pages 265-274. | 1-30 |

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 27 JANUARY 1997 | 19 FEB 1997 |
| Name and mailing address of the ISA/US<br>Commissioner of Patents and Trademarks<br>Box PCT<br>Washington, D.C. 20231 | Authorized officer<br>SCOTT D. PRIEBE |
| Facsimile No. (703) 305-3230 | Telephone No. (703) 308-0196 |

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/16342

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A | WO 95/21944 A1 (SMITHKLINE BEECHAM CORPORATION) 17 August 1995, page 4, lines 1-4, page 5, lines 31-37, page 17, lines 15-27, page 18, lines 30-35, page 20, line 23 to page 21, line 4. | 1-30 |